Updating Probabilities

Ariel Caticha and Adom Giffin

Department of Physics, University at Albany–SUNY, Albany, NY 12222,USA

Abstract. We show that Skilling's method of induction leads to a unique general theory of inductive inference, the method of Maximum relative Entropy (ME). The main tool for updating probabilities is the logarithmic relative entropy; other entropies such as those of Renyi or Tsallis are ruled out. We also show that Bayes updating is a special case of ME updating and thus, that the two are completely compatible.

INTRODUCTION

The method of Maximum (relative) Entropy (ME) [1, 2, 3] is designed for updating probabilities when new information is given in the form of a constraint on the family of allowed posteriors. This is in contrast with the older MaxEnt method [4] which was designed to assign rather than update probabilities. The objective of this paper is to strengthen the ME method in two ways.

In [3] the axioms that define the ME method have been distilled down to three. In this work the justification of the method is improved by considerably weakening the axiom that deals with independent subsystems. We adopt a consistency axiom similar to that proposed by Shore and Johnson [1]: When two systems are independent it should not matter whether the inference procedure treats them separately or jointly. The merit of such a consistency axiom is that it is very compelling. Nevertheless, the mathematical implementation of the axiom has been criticized by Karbelkar [5] and by Uffink [6]. In their view it fails to single out the usual logarithmic entropy as the unique tool for updating. It merely restricts the form of the entropy to a one-dimensional continuum labeled by a parameter η. The resulting η-entropies are equivalent to those proposed by Renyi [7] and by Tsallis [8] in the sense that they update probabilities in the same way.

The main result of this paper is to go beyond the insights of Karlbelkar and Uffink, and show that the consistency axiom selects a unique, universal value for the parameter η and this value ($\eta = 0$) corresponds to the usual logarithmic entropy. The advantage of our approach is that it shows precisely how it is that η-entropies with $\eta \neq 0$ are ruled out as tools for updating.

Our second objective is mostly pedagogical. The preeminent updating method is based on Bayes' rule and we want to discuss its relation with the ME method. We start by drawing a distinction between Bayes' *theorem*, which is a straightforward consequence of the product rule for probabilities, and Bayes' *rule*, which is the actual updating rule. We show that Bayes' rule can be derived as a special case of the ME method, a result that was first obtained by Williams [11, 12] long before the logical status of the ME method had been sufficiently clarified. The virtue of our derivation, which hinges on

translating information in the form of data into constraints that can be processed using ME, is that it is particularly clear. It throws light on Bayes' rule and demonstrates its complete compatibility with ME updating. A slight generalization of the same ideas shows that Jeffrey's updating rule is also a special case of the ME method.

ENTROPY AS A TOOL FOR UPDATING PROBABILITIES

Our objective is to devise a general method to update from a prior distribution $q(x)$ to a posterior distribution $p(x)$ when new information becomes available. By information, in its most general form, we mean a set of constraints on the family of acceptable posterior distributions. *Information is whatever constrains our beliefs*.

To carry out the update we proceed by ranking the allowed probability distributions according to increasing *preference*. This immediately raises two questions: (a) how is the ranking implemented and (b) what makes one distribution preferable over another? The answer to (a) is that any useful ranking scheme must be transitive (if P_1 is better than P_2, and P_2 is better than P_3, then P_1 is better than P_3), and therefore it can be implemented by assigning a real number $S[P]$ to each P in such a way that if P_1 is preferred over P_2, then $S[P_1] > S[P_2]$. The preferred P is that which maximizes the "entropy" $S[P]$. This explains why entropies are real numbers and why they are meant to be maximized.

Question (b), the criterion for preference, is implicitly answered once the functional form of the entropy $S[P]$ that defines the ranking scheme is chosen. The basic strategy is inductive. We follow Skilling's method of induction [2]: (1) If an entropy $S[P]$ of universal applicability exists, it must apply to special examples. (2) If in a certain example the best distribution is known, then this knowledge constrains the form of $S[P]$. Finally, (3) if enough examples are known, then $S[P]$ will be completely determined. (Of course, the known examples might turn out to be incompatible with each other, in which case there is no universal $S[P]$ that accommodates them all.)

It is perhaps worth emphasizing that in this approach entropy is a tool for reasoning which requires no interpretation in terms of heat, multiplicities, disorder, uncertainty, or amount of information. *Entropy needs no interpretation*. We do not need to know what it means, we only need to know how to use it.

The known special examples, which are called the "axioms" of ME, reflect the conviction that what was learned in the past is important and should not be easily ignored. The chosen posterior distribution should coincide with the prior as closely as possible and one should only update those aspects of one's beliefs for which corrective new evidence has been supplied. The first two axioms are listed below. (The motivation and detailed proofs are found in [3].)

Axiom 1: Locality. *Local information has local effects.*
When the new information does not refer to a domain D of the variable x the conditional probabilities $p(x|D)$ need not be revised. The consequence of the axiom is that non-overlapping domains of x contribute additively to the entropy: $S[P] = \int dx F(P(x),x)$ where F is some unknown function.

Axiom 2: Coordinate invariance. *The ranking should not depend on the system of coordinates.*

The coordinates that label the points x are arbitrary; they carry no information. The consequence of this axiom is that $S[P] = \int dx\, m(x)\Phi(P(x)/m(x))$ involves coordinate invariants such as $dx\, m(x)$ and $P(x)/m(x)$, where the functions $m(x)$ (which is a density) and Φ are, at this point, still undetermined.

Next we make a second use of the locality axiom and allow domain D to extend over the whole space. Axiom 1 then asserts that *when there is no new information there is no reason to change one's mind.* When there are no constraints the selected posterior distribution should coincide with the prior distribution. This eliminates the arbitrariness in the density $m(x)$: up to normalization $m(x)$ is the prior distribution, $m(x) \propto q(x)$.

In [3] the remaining unknown function Φ was determined using the following axiom:
Old Axiom 3: Subsystem independence. *When a system is composed of subsystems that are* ***believed*** *to be independent it should not matter whether the inference procedure treats them separately or jointly.*
Let us be very explicit about what this axiom means. Consider a system composed of two subsystems which our prior evidence has led us to believe are independent. This belief is reflected in the prior distribution: if the subsystem priors are $q_1(x_1)$ and $q_2(x_2)$, then the prior for the whole system is the product $q_1(x_1)q_2(x_2)$. Further suppose that new information is acquired such that $q_1(x_1)$ is updated to $p_1(x_1)$ and that $q_2(x_2)$ is updated to $p_2(x_2)$. Nothing in this new information requires us to revise our previous assessment of independence, therefore there is no need to change our minds, and the function Φ must be such that the prior for the whole system $q_1(x_1)q_2(x_2)$ should be updated to $p_1(x_1)p_2(x_2)$.

This idea is implemented as follows: First we treat the two subsystems separately. Suppose that for subsystem 1 maximizing

$$S_1[P_1,q_1] = \int dx_1\, q_1(x_1)\Phi\left(\frac{P_1(x_1)}{q_1(x_1)}\right), \tag{1}$$

subject to constraints $\mathscr{C}_1$ on the marginal distribution $P_1(x_1) = \int dx_2\, P(x_1,x_2)$ selects the posterior $p_1(x_1)$. The constraints $\mathscr{C}_1$ could, for example, include normalization, or they could involve the known expected value of a function $f_1(x_1)$,

$$\int dx_1 f_1(x_1)P_1(x_1) = \int dx_1 dx_2\, f_1(x_1)P(x_1,x_2) = F_1\ . \tag{2}$$

Similarly, suppose that for subsystem 2 maximizing the corresponding $S_2[P_2,q_2]$ subject to constraints $\mathscr{C}_2$ on $P_2(x_2) = \int dx_1\, P(x_1,x_2)$ selects the posterior $p_2(x_2)$.

Next we treat the subsystems jointly and maximize the joint entropy,

$$S[P,q_1q_2] = \int dx_1 dx_2\, q_1(x_1)q_2(x_2)\Phi\left(\frac{P(x_1,x_2)}{q_1(x_1)q_2(x_2)}\right), \tag{3}$$

subject to the *precisely the same constraints* on the joint distribution P. The function Φ is determined by the requirement that the selected posterior be p_1p_2. As shown in [3] this leads to the logarithmic form

$$S[P,q] = -\int dx\, P(x)\log\frac{P(x)}{q(x)}\ . \tag{4}$$

THE NEW INDEPENDENCE AXIOM

Next we replace our old axiom 3 by an axiom which is more convincing axiom because it is an explicit requirement of consistency.

New Axiom 3: Consistency for independent subsystems. *When a system is composed of subsystems that are* ***known*** *to be independent it should not matter whether the inference procedure treats them separately or jointly.*

Again, we have to be very explicit about what this axiom means and how it differs from the old one. When the subsystems are treated separately the inference proceeds exactly as described before: for subsystem 1 maximize the entropy $S_1[P_1, q_1]$ subject to the constraints $\mathscr{C}_1$ to select a posterior p_1 and similarly for subsystem 2 to select p_2. The important difference is introduced when the subsystems are treated jointly. Since we are only concerned with those special examples where we *know* that the subsystems are independent, we are *required* to search for the posterior within the restricted family of joint distributions that take the form of a product $P = P_1P_2$; this is an *additional* constraint over and above the original $\mathscr{C}_1$ and $\mathscr{C}_2$.

In the previous case we chose Φ so as to maintain independence because there was no evidence against it. Here we impose independence by hand as an additional constraint for the stronger reason that the subsystems are known to be independent. At first sight it appears that the new axiom does not place as stringent a restriction on the general form of Φ: it would seem that Φ has been relieved of its responsibility of enforcing independence because it is up to us to impose it explicitly by hand. However, as we shall see, the fact that we seek an entropy S of *general* applicability and that we require consistency for *all possible* independent subsystems is sufficiently restrictive.

The new constraint $P = P_1P_2$ is easily implemented by direct substitution. Instead of maximizing the joint entropy, $S[P, q_1q_2]$, we now maximize

$$S[P_1P_2, q_1q_2] = \int dx_1dx_2\, q_1(x_1)q_2(x_2)\Phi\left(\frac{P_1(x_1)P_2(x_2)}{q_1(x_1)q_2(x_2)}\right), \tag{5}$$

under independent variations δP_1 and δP_2 subject to the same constraints $\mathscr{C}_1$ and $\mathscr{C}_2$ and we choose Φ by imposing that the updating leads to the posterior $p_1(x_1)p_2(x_2)$.

Consistency for identical independent subsystems

Here we show that applying the axiom to subsystems that happen to be identical restricts the entropy functional to a member of the one-parameter family given by

$$S_\eta[P, q] = -\int dx P(x)\left(\frac{P(x)}{q(x)}\right)^\eta \quad \text{for} \quad \eta \neq -1, 0\,. \tag{6}$$

Since entropies that differ by additive or multiplicative constants are equivalent in that they induce the same ranking scheme, we could equally well have written

$$S_\eta[P, q] = \frac{1}{\eta(\eta+1)}\left(1 - \int dx P^{\eta+1}q^{-\eta}\right)\,. \tag{7}$$

This is convenient because the entropies for $\eta = 0$ and $\eta = -1$ can be obtained by taking the appropriate limits. For $\eta \to 0$ use $y^\eta = \exp \eta \log y \approx 1 + \eta \log y$ to obtain the usual logarithmic entropy, $S_0[P,q] = S[P,q]$ in eq.(4). Similarly, for $\eta \to -1$ we get $S_{-1}[P,q] = S[q,P]$.

The proof below is based upon and extends a previous proof by Karbelkar [5]. He showed that belonging to the family of η-entropies is a sufficient condition to satisfy the consistency axiom for identical systems and he conjectured but did not prove that this was perhaps also a necessary condition. Although necessity was not essential to his argument it is crucial for ours. We show below that for identical subsystems there are no acceptable entropies outside this family.

Proof

First we treat the subsystems separately. For subsystem 1 we maximize the entropy $S_1[P_1,q_1]$ subject to normalization and the constraint $\mathscr{C}_1$ in eq.(2). Introduce Lagrange multipliers α_1 and λ_1,

$$\delta \left[S_1[P_1,q_1] - \lambda_1 \left(\int dx_1 f_1 P_1 - F_1 \right) - \alpha_1 \left(\int dx_1 P_1 - 1 \right) \right] = 0, \tag{8}$$

which gives

$$\Phi' \left(\frac{P_1(x_1)}{q_1(x_1)} \right) = \lambda_1 f_1(x_1) + \alpha_1 \ , \tag{9}$$

where the prime indicates a derivative with respect to the argument, $\Phi'(y) = d\Phi(y)/dy$. For subsystem 2 we need only consider the extreme situation where the constraints $\mathscr{C}_2$ determine the posterior completely: $P_2(x_2) = p_2(x_2)$.

Next we treat the subsystems jointly. The constraints $\mathscr{C}_2$ are easily implemented by direct substitution and thus, we maximize the entropy $S[P_1 p_2, q_1 q_2]$ by varying over P_1 subject to normalization and the constraint $\mathscr{C}_1$ in eq.(2). Introduce Lagrange multipliers α and λ,

$$\delta \left[S[P_1 p_2, q_1 q_2] - \lambda \left(\int dx_1 f_1 P_1 - F_1 \right) - \alpha \left(\int dx_1 P_1 - 1 \right) \right] = 0, \tag{10}$$

which gives

$$\int dx_2 \, p_2 \Phi' \left(\frac{P_1 p_2}{q_1 q_2} \right) = \lambda[p_2] f_1(x_1) + \alpha[p_2] \ , \tag{11}$$

where the multipliers λ and α are independent of x_1 but could in principle be functionals of p_2.

The consistency condition that constrains the form of Φ is that if the solution to eq.(9) is $p_1(x_1)$ then the solution to eq.(11) must also be $p_1(x_1)$, and this must be true irrespective of the choice of $p_2(x_2)$. Let us then consider a small change $p_2 \to p_2 + \delta p_2$ that preserves the normalization of p_2. First introduce a Lagrange multiplier α_2 and

rewrite eq.(11) as

$$\int dx_2\, p_2 \Phi'\left(\frac{p_1 p_2}{q_1 q_2}\right) - \alpha_2 \left[\int dx_2\, p_2 - 1\right] = \lambda[p_2] f_1(x_1) + \alpha[p_2]\ , \tag{12}$$

where we have replaced P_1 by the known solution p_1 and thereby effectively transformed eqs.(9) and (11) into an equation for Φ. The $\delta p_2(x_2)$ variation gives,

$$\Phi'\left(\frac{p_1 p_2}{q_1 q_2}\right) + \frac{p_1 p_2}{q_1 q_2}\Phi''\left(\frac{p_1 p_2}{q_1 q_2}\right) = \frac{\delta\lambda}{\delta p_2} f_1(x_1) + \frac{\delta\alpha}{\delta p_2} + \alpha_2\ . \tag{13}$$

Next use eq.(9) to eliminate $f_1(x_1)$,

$$\Phi'\left(\frac{p_1 p_2}{q_1 q_2}\right) + \frac{p_1 p_2}{q_1 q_2}\Phi''\left(\frac{p_1 p_2}{q_1 q_2}\right) = A[\frac{p_2}{q_2}]\Phi'\left(\frac{p_1}{q_1}\right) + B[\frac{p_2}{q_2}]\ , \tag{14}$$

where

$$A[\frac{p_2}{q_2}] = \frac{1}{\lambda_1}\frac{\delta\lambda}{\delta p_2} \quad \text{and} \quad B[\frac{p_2}{q_2}] = -\frac{\delta\lambda}{\delta p_2}\frac{\alpha_1}{\lambda_1} + \frac{\delta\alpha}{\delta p_2} + \alpha_2\ , \tag{15}$$

are at this point unknown functionals of p_2/q_2. Differentiating eq.(14) with respect to x_1 the B term drops out and we get

$$A[\frac{p_2}{q_2}] = \left[\frac{d}{dx_1}\Phi'\left(\frac{p_1}{q_1}\right)\right]^{-1} \frac{d}{dx_1}\left[\Phi'\left(\frac{p_1 p_2}{q_1 q_2}\right) + \frac{p_1 p_2}{q_1 q_2}\Phi''\left(\frac{p_1 p_2}{q_1 q_2}\right)\right]\ , \tag{16}$$

which shows that A is not a functional but a mere function of p_2/q_2. Substituting back into eq.(14) we see that the same is true for B. Therefore eq.(14) can be written as

$$\Phi'(y_1 y_2) + y_1 y_2 \Phi''(y_1 y_2) = A(y_2)\Phi'(y_1) + B(y_2)\ , \tag{17}$$

where $y_1 = p_1/q_1$, $y_2 = p_2/q_2$, and $A(y_2)$, $B(y_2)$ are unknown functions of y_2. If we specialize to identical subsystems for which we can exchange the labels $1 \leftrightarrow 2$, we get

$$A(y_2)\Phi'(y_1) + B(y_2) = A(y_1)\Phi'(y_2) + B(y_1)\ . \tag{18}$$

To find the unknown functions A and B differentiate with respect to y_2,

$$A'(y_2)\Phi'(y_1) + B'(y_2) = A(y_1)\Phi''(y_2) \tag{19}$$

and then with respect to y_1 to get

$$\frac{A'(y_1)}{\Phi''(y_1)} = \frac{A'(y_2)}{\Phi''(y_2)} = a = \text{const}\ . \tag{20}$$

Integrating,

$$A(y_1) = a\Phi'(y_1) + b\ . \tag{21}$$

Substituting back into eq.(19) and integrating gives

$$B'(y_2) = b\Phi''(y_2) \quad \text{and} \quad B(y_2) = b\Phi'(y_2) + c \ , \tag{22}$$

where b and c are constants. We can check that $A(y)$ and $B(y)$ are indeed solutions of eq.(18). Substituting into eq.(17) gives

$$\Phi'(y_1y_2) + y_1y_2\Phi''(y_1y_2) = a\Phi'(y_1)\Phi'(y_2) + b\left[\Phi'(y_1) + \Phi'(y_2)\right] + c \ . \tag{23}$$

This is a peculiar differential equation. We can think of it as one differential equation for $\Phi'(y_1)$ for each given constant value of y_2 but there is a complication in that the various (constant) coefficients $\Phi'(y_2)$ are themselves unknown. To solve for Φ choose a fixed value of y_2, say $y_2 = 1$,

$$y\Phi''(y) - \eta\Phi'(y) - \kappa = 0 \ , \tag{24}$$

where $\eta = a\Phi'(1) + b - 1$ and $\kappa = b\Phi'(1) + c$. To eliminate the constant κ differentiate with respect to y,

$$y\Phi''' + (1-\eta)\Phi'' = 0 \ , \tag{25}$$

which is a linear homogeneous equation and is easy to integrate. For a generic value of η the solution is

$$\Phi''(y) \propto y^{\eta-1} \Rightarrow \Phi'(y) = \alpha y^{\eta} + \beta \ . \tag{26}$$

The constants α and β are chosen so that this is a solution of eq.(23) for all values of y_2 (and not just for $y_2 = 1$). Substituting into eq.(23) and equating the coefficients of various powers of y_1y_2, y_1, and y_2 gives three conditions on the two constants α and β,

$$\alpha(1+\eta) = a\alpha^2, \quad 0 = a\alpha\beta + b\alpha, \quad \beta = a\beta^2 + 2b\beta + c \ . \tag{27}$$

The nontrivial ($\alpha \neq 0$) solutions are $\alpha = (1+\eta)/a$ and $\beta = -b/a$, while the third equation gives $c = b(1-b)/4a$. We conclude that for generic values of η the solution of eq.(23) is

$$\Phi(y) = \frac{1}{a}y^{\eta+1} - \frac{b}{a}y + C \ , \tag{28}$$

where C is a new constant. Choosing $a = -\eta(\eta+1)$ and $b = 1 + Ca$ we obtain eq.(7).

For the special values $\eta = 0$ and $\eta = -1$ one can either first take the limit of the differential eq.(25) and then find the relevant solutions, or one can first solve the differential equation for general η and then take the limit of the solution eq.(7) as described earlier. Either way one obtains (up to additive and multiplicative constants which have no effect on the ranking scheme) the entropies $S_0[P,q] = S[P,q]$ and $S_{-1}[P,q] = S[q,P]$.

Consistency for non-identical subsystems

Let us summarize our results so far. The goal is to update probabilities by ranking the distributions according to an entropy S that is of general applicability. The functional form of the entropy S has been constrained down to a member of the one-dimensional

family S_η. One might be tempted to conclude (see [5, 6]) that there is no S of universal applicability; that inferences about different systems ought to be carried out with different η-entropies. But we have not yet exhausted the full power of our new axiom 3.

To proceed further we ask: What is η? Is it a property of the individual carrying out the inference or of the system under investigation? The former makes no sense; we insist that the updating must be objective in that different individuals with the same prior and the same information must make the same inference. Therefore the "inference parameter" η must be a characteristic of the system.

Consider two different systems characterized by η_1 and η_2. Let us further suppose that these systems are independent (perhaps system 1 is here on Earth while the other lives in a distant galaxy) so that they fall under the jurisdiction of the new axiom 3; inferences about system 1 are carried out with $S_{\eta_1}[P_1, q_1]$ while inferences about system 2 require $S_{\eta_2}[P_2, q_2]$. For the combined system we are also required to use an η-entropy $S_\eta[P_1P_2, q_1q_2]$. The question is what η do we choose that will lead to consistent inferences whether we treat the systems separately or jointly. The results of the previous section indicate that a joint inference with $S_\eta[P_1P_2, q_1q_2]$ is equivalent to separate inferences with $S_\eta[P_1, q_1]$ and $S_\eta[P_2, q_2]$. Therefore we must choose $\eta = \eta_1$ and also $\eta = \eta_2$ which is possible only when $\eta_1 = \eta_2$. But this is not all: any other system whether here on Earth or elsewhere that happens to be independent of the distant system 2 must also be characterized by the same inference parameter $\eta = \eta_2 = \eta_1$ even if it is correlated with system 1. Thus all systems have the same η whether they are independent or not.

The power of a consistency argument resides in its universal applicability: if a general expression for $S[P, q]$ exists then it must be of the form $S_\eta[P, q]$ where η is a universal constant. The remaining problem is to determine this universal η. One possibility is to determine η experimentally: are there systems for which inferences based on a known value of η have repeatedly led to success? The answer is yes; they are quite common.

The next step in our argument is provided by the work of Jaynes [4] who showed that statistical mechanics and thus thermodynamics are theories of inference based on the value $\eta = 0$. His method, called MaxEnt, can be interpreted as the special case of the ME when one updates from a uniform prior using the Gibbs-Shannon entropy. Thus, it is an experimental fact without any known exceptions that inferences about *all* physical, chemical and biological systems that are in thermal equilibrium or close to it can be carried out by assuming that $\eta = 0$. Let us emphasize that this is not an obscure and rare example of purely academic interest; these systems comprise essentially all of natural science. (Included is every instance where it is useful to introduce a notion of temperature.)

In conclusion: consistency for non-identical systems requires that η be a universal constant and there is abundant experimental evidence for its value being $\eta = 0$. Other η-entropies may be useful for other purposes but the logarithmic entropy $S[P, q]$ in eq.(4) provides the only consistent ranking criterion for updating probabilities that can claim general applicability.

BAYES UPDATING

The two preeminent updating methods are the ME method discussed above and Bayes' rule. The choice between the two methods has traditionally been dictated by the nature of the information being processed (either constraints or observed data) but questions about their compatibility are regularly raised. Our goal here is to show that these two updating strategies are completely consistent with each other. Let us start by drawing a distinction between Bayes' theorem and Bayes' rule.

Bayes' theorem and Bayes' rule

The goal here is to update our beliefs about the values of one or several quantities $\theta \in \Theta$ on the basis of observed values of variables $x \in \mathcal{X}$ and of the known relation between them represented by a specific model. The first important point to make is that attention must be focused on the joint distribution $P_{\rm old}(x,\theta)$. Indeed, being a consequence of the product rule, Bayes' theorem requires that $P_{\rm old}(x,\theta)$ be defined and that assertions such as "x *and* θ" be meaningful; the relevant space is neither $\mathcal{X}$ nor Θ but the product $\mathcal{X} \times \Theta$. The label "old" is important. It has been attached to the joint distribution $P_{\rm old}(x,\theta)$ because this distribution codifies our beliefs about x and about θ before the information contained in the actual data has been processed. The standard derivation of Bayes' theorem invokes the product rule,

$$P_{\rm old}(x,\theta) = P_{\rm old}(x)P_{\rm old}(\theta|x) = P_{\rm old}(\theta)P_{\rm old}(x|\theta) \ , \tag{29}$$

so that

$$P_{\rm old}(\theta|x) = P_{\rm old}(\theta)\frac{P_{\rm old}(x|\theta)}{P_{\rm old}(x)} \ . \qquad \text{(Bayes' theorem)}$$

It is important to realize that at this point there has been no updating. Our beliefs have not changed. All we have done is rewrite what we knew all along in $P_{\rm old}(x,\theta)$. Bayes' *theorem* is an identity that follows from requirements on how we should consistently assign degrees of belief. Whether the justification of the product rule is sought through Cox's consistency requirement and regraduation or through a Dutch book betting coherence argument, the theorem is valid irrespective of whatever data will be or has been collected. Our notation, with the label "old" throughout, makes this point explicit.

The real updating from the old prior distribution $P_{\rm old}(\theta)$ to a new posterior distribution $P_{\rm new}(\theta)$ occurs when we take into account the values of x that have actually been observed, which we will denote with a capital X. This requires a new assumption and the natural choice is that the updated distribution $P_{\rm new}(\theta)$ be given by Bayes' *rule*,

$$P_{\rm new}(\theta) = P_{\rm old}(\theta|X) \ . \qquad \text{(Bayes rule)}$$

Combining Bayes' theorem with Bayes' rule leads to the standard equation for Bayes updating,

$$P_{\rm new}(\theta) = P_{\rm old}(\theta)\frac{P_{\rm old}(X|\theta)}{P_{\rm old}(X)} \ . \tag{30}$$

The assumption embodied in Bayes' rule is extremely reasonable: we maintain those old beliefs about θ that are consistent with data values that have turned out to be true. Data values that were not observed are discarded because they are now known to be false.

This argument is indeed so compelling that it may seem unnecessary to seek any further justification for the Bayes' rule assumption. However, we deal here with such a basic algorithm for information processing – it is fundamental to all experimental science – that even such a self-evident assumption should be carefully examined and its compatibility with the ME method should be verified.

Bayes' rule from ME

Our first concern when using the ME method to update from a prior to a posterior distribution is to define the space in which the search for the posterior will be conducted. We argued above that the relevant space is the product $\mathscr{X} \times \Theta$. Therefore the selected posterior $P_{\rm new}(x,\theta)$ is that which maximizes

$$S[P,P_{\rm old}] = -\int dx d\theta\ P(x,\theta) \log \frac{P(x,\theta)}{P_{\rm old}(x,\theta)} \tag{31}$$

subject to the appropriate constraints.

Next, the information being processed, the observed data X, must be expressed in the form of a constraint on the allowed posteriors. Clearly, the family of posteriors that reflects the fact that x is now known to be X is such that

$$P(x) = \int d\theta\ P(x,\theta) = \delta(x-X)\ . \tag{32}$$

This amounts to an *infinite* number of constraints: there is one constraint on $P(x,\theta)$ for each value of the variable x and each constraint will require its own Lagrange multiplier $\lambda(x)$. Furthermore, we impose the usual normalization constraint,

$$\int dx d\theta\ P(x,\theta) = 1\ . \tag{33}$$

Maximize S subject to these constraints,

$$\delta\left\{S + \int dx\, \lambda(x) \left[\int d\theta\ P(x,\theta) - \delta(x-X)\right] + \alpha\left[\int dx d\theta\ P(x,\theta) - 1\right]\right\} = 0\ , \tag{34}$$

and the selected posterior is

$$P_{\rm new}(x,\theta) = P_{\rm old}(x,\theta)\, \frac{e^{\lambda(x)}}{Z}\ , \tag{35}$$

where the normalization Z is

$$Z = e^{-\alpha+1} = \int dx d\theta\, P_{\rm old}(x,\theta)\, e^{\lambda(x)}\ , \tag{36}$$

and the multipliers $\lambda(x)$ are determined from eq.(32),

$$\int d\theta\ P_{\rm old}(x,\theta) \frac{e^{\lambda(x)}}{Z} = P_{\rm old}(x) \frac{e^{\lambda(x)}}{Z} = \delta(x-X)\ . \tag{37}$$

Therefore, substituting $e^{\lambda(x)}$ back into eq.(35),

$$P_{\rm new}(x,\theta) = \frac{P_{\rm old}(x,\theta)\,\delta(x-X)}{P_{\rm old}(x)} = \delta(x-X)P_{\rm old}(\theta|x)\ . \tag{38}$$

The new marginal distribution for θ is

$$P_{\rm new}(\theta) = \int dx P_{\rm new}(x,\theta) = P_{\rm old}(\theta|X)\ , \tag{39}$$

which is Bayes' rule! Bayes updating is a special case of ME updating.

To summarize: the prior $P_{\rm old}(x,\theta) = P_{\rm old}(x)P_{\rm old}(\theta|x)$ is updated to the posterior $P_{\rm new}(x,\theta) = P_{\rm new}(x)P_{\rm new}(\theta|x)$ where $P_{\rm new}(x) = \delta(x-X)$ is fixed by the observed data while $P_{\rm new}(\theta|x) = P_{\rm old}(\theta|x)$ remains unchanged. Note that in accordance with the philosophy that drives the ME method *one only updates those aspects of one's beliefs for which corrective new evidence has been supplied*.

The generalization to situations where there is some uncertainty about the actual data is straightforward. In this case the marginal $P(x)$ in eq.(32) is not a δ function but a known distribution $P_D(x)$. The selected posterior $P_{\rm new}(x,\theta) = P_{\rm new}(x)P_{\rm new}(\theta|x)$ is easily shown to be $P_{\rm new}(x) = P_D(x)$ with $P_{\rm new}(\theta|x) = P_{\rm old}(\theta|x)$ remaining unchanged. This leads to Jeffrey's conditionalization rule,

$$P_{\rm new}(\theta) = \int dx\, P_{\rm new}(x,\theta) = \int dx\, P_D(x)P_{\rm old}(\theta|x)\ . \tag{40}$$

CONCLUSIONS

We have shown that Skilling's method of induction has led to a unique general theory of inductive inference, the ME method. The whole approach is extremely conservative. First, the axioms merely instruct us what not to update – do not change your mind except when forced by new information. Second, the validity of the method does not depend on any particular interpretation of the notion of entropy – entropy needs no interpretation.

Our derivation of the consequences of the new axiom show that when applied to identical subsystems they restrict the entropy to a member of the η-entropy family. Its further application to non-identical systems shows that consistency requires that η be a universal constant which must take the value $\eta = 0$ in order to account for the empirical success of the inference theory we know as statistical mechanics. Thus, the unique tool for updating probabilities is the logarithmic relative entropy. Other entropies with $\eta \neq 0$ or those of Renyi or Tsallis are ruled out; they may be useful for other purposes but not for inference.

Finally we explored the compatibility of Bayes and ME updating. After pointing out the distinction between Bayes' theorem and the Bayes' updating rule, we showed that Bayes' rule is a special case of ME updating by translating information in the form of data into constraints that can be processed using ME.

Acknowledgements: We would like to acknowledge valuable discussions with N. Caticha, R. Fischer, M. Grendar, K. Knuth, C. Rodríguez, and A. Solana-Ortega.

REFERENCES

1. J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **IT-26**, 26 (1980); IEEE Trans. Inf. Theory **IT-27**, 26 (1981).
2. J. Skilling, "The Axioms of Maximum Entropy" in *Maximum-Entropy and Bayesian Methods in Science and Engineering*, G. J. Erickson and C. R. Smith (eds.) (Kluwer, Dordrecht, 1988).
3. A. Caticha, "Relative Entropy and Inductive Inference," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by G. Erickson and Y. Zhai, AIP Conf. Proc. **707**, 75 (2004) (arXiv.org/abs/physics/0311093).
4. E. T. Jaynes, Phys. Rev. **106**, 620 and **108**, 171 (1957); R. D. Rosenkrantz (ed.), *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics* (Reidel, Dordrecht, 1983); E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).
5. S. N. Karbelkar, Pramana – J. Phys. **26**, 301 (1986).
6. J. Uffink, Stud. Hist. Phil. Mod. Phys. **26B**, 223 (1995).
7. A. Renyi, "On measures of entropy and information," *Proc. 4th Berkeley Simposium on Mathematical Statistics and Probability*, Vol 1, p. 547-461 (U. of California Press, 1961).
8. C. Tsallis, J. Stat. Phys. **52**, 479 (1988); "Nonextensive statistical mechanics: a brief review of its present status," online at arXiv.org/abs/cond-mat/0205571. Critiques of Tsallis' non-extensive entropy are given in [9]; derivations of Tsallis' distributions from standard principles of statistical mechanics are given in [10].
9. B. La Cour and W. C. Schieve, Phys. Rev. E **62**, 7494 (2000); M. Nauenberg, Phys. Rev. E **67**, 036114 (2003).
10. A. R. Plastino and A. Plastino, Phys. Lett. **A193**, 140 (1994); G. Wilk and Z. Wlodarczyk, Phys. Rev. Lett. **84**, 2770 (2000).
11. P. M. Williams, Brit. J. Phil. Sci. **31**, 131 (1980).
12. P. Diaconis and S. L. Zabell, J. Am. Stat. Assoc. **77**, 822 (1982).