entropy

MDPI

Article

A Definition of Conditional Probability with Non-Stochastic Information

Pier Giovanni Bissiri [1,*,†] and Stephen G. Walker [2,†]

[1] School of Mathematics, Statistics and Physics, Herschel Building, Newcastle University, Newcastle upon Tyne NE1 7RU, UK
[2] Department of Mathematics, University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA; s.g.walker@math.utexas.edu
* Correspondence: pier-giovanni.bissiri@newcastle.ac.uk; Tel.: +44-(0)-191-208-8327
† These authors contributed equally to this work.

Received: 21 June 2018; Accepted: 31 July 2018; Published: 3 August 2018

Abstract: The current definition of a conditional probability enables one to update probabilities only on the basis of stochastic information. This paper provides a definition for conditional probability with non-stochastic information. The definition is derived by a set of axioms, where the information is connected to the outcome of interest via a loss function. An illustration is presented.

Keywords: conditional probability; loss function; non-stochastic information

1. Introduction

The theory of conditional probability is a well-established mathematical theory that provides a procedure to update probabilities by taking new information into account. The new work in this paper is motivated by the fact that such a procedure is available only if the information which is used to update the probability concerns stochastic events, that is, events to which a probability is assigned. In other words, such information needs to be already included in the probability model. The statistical implications are most pertinent to Bayesian inference where non-stochastic information, obtained by experts for example, is often employed.

To set the scene, Ω denotes the set of possible outcomes of an experiment, which we assume is finite, $p(\omega)$ is a probability mass function on Ω, and I is an arbitrary piece of information about the unknown outcome of the experiment, i.e., information for which one would consider revising $p(\omega)$. We argue that there *must* be a formal way to derive the updated probability mass function, which we write as $p_I(\omega)$. For example, Ω might represent a horse race and a bookmaker has set p. It starts to rain, so $I =$ "it is raining", and the bookmaker revises p to p_I. One imagines that the bookmaker does so from experience and uses no formal rule to implement the change. This lack of a formal rule then questions the extent of the relevance of subjective probability if one can appeal to no formal mechanism for revising belief distributions with arbitrary, yet relevant, information. Recall that the formal Bayes conditional probability rule requires the specification of $p(I|\omega)$ for all possible pieces of information I that one could receive–to us, at least, this is an impossible task.

It is impossible to set the probabilities $p(I|\omega)$ *a priori* for all possible I, yet this is required for the validation of the mathematical implementation of the Bayes rule. However, setting a loss function $l(\omega, I)$ once I has been received, is, we argue, a totally viable task. This would establish the loss, on some scale, e.g., monetary, if the elementary event ω occurs once I has been received. Our formal updating rule relies on the construction of such $l(\omega, I)$.

I is a piece of information for which it is possible to construct a loss function $l(\cdot, I)$ on Ω that is valued into $[0, \infty]$ and is not identically infinite. So, $l(\omega, I)$ is the loss that the experimenter assigns

to the outcome ω in Ω when holding the piece of information I. We look at an illustration involving such $l(\omega, I)$ later on in the paper. Once I has been received, the probability of updating is P, where $P(B) = \sum_{\omega \in B} p(\omega)$ for every $B \subset \Omega$.

For example, if I is the information "event $B \subset \Omega$ has occurred" and

$$l(\omega, I) = \begin{cases} \infty & \omega \notin B \\ \\ 0 & \omega \in B \end{cases} \tag{1}$$

then our formal rule will coincide with the usual definition of conditional probability.

When the standard definition of conditional probability does not apply, e.g., I is not such a statement or B is not a subset of Ω, for example, an alternative definition based on a mathematical decision theoretic framework can be used. When information received is non-stochastic, but relevant to an outcome of interest, we cannot use a probability distribution and so we need an alternative way to connect the information I with the outcome of interest ω. We do so using loss functions and a set of axioms.

This paper provides a definition for conditional probability on the basis of I. Reference [1] addressed this issue considering the minimization of a cumulative loss function which involved g-divergences. In this paper, instead, the definition of conditional probability is solely based on a set of axioms. This idea was developed by reference [2] who defined a framework for general Bayesian inference and discussed its application to important statistical problems. The present paper also addresses the issue of calibrating the conditional probability with non-stochastic information.

1.1. Relationship to the Literature

In the literature, alternative definitions of conditional probability, such as the Jeffrey's Rule of conditioning, are given where new information is not put in terms of the occurrence of an event included in the model. These definitions rely on the assumption that information can be given in the form of a constraint on the probability. Constraints considered are of the type

$$\sum_{\omega \in \Omega} g(\omega) p_I(\omega) > 0, \tag{2}$$

where g is a measurable real function on Ω, and p_I is the updated probability function. The idea is to minimize the Kullback–Leibler divergence subject to the constraint (2), which represents the information I. This problem can be solved using Lagrange multipliers.

For more detail about conditionalization based upon constraints on the conditional distribution, see references [3–8]. Our approach is different as we can deal with more types of information. On the other hand, our definition can encompass potentially arbitrary information about the outcome; all we need is to construct a loss function $l(\omega, I)$ for each ω in Ω once I has been received.

1.2. Motivation

The need for a definition of conditional probability outside of the usual set-up is the avoidance of paradoxes where one knows event B has occured yet is not a subset of Ω. Paradoxes arise when B is not deemed to be part of the stochastic model yet is subsequently forced to be so.

Such difficulties arise in different puzzles, such as, for instance, Freund's puzzle of the two aces, introduced by reference [9]. For other puzzles about conditional probability, see, for instance, reference Gardner [10].

These puzzles have been widely used to discuss the concept of conditional probability. According to reference [11], such a concept is justifiable only on the basis of "a set of rules that tell, at each step, what can happen next", which he calls "protocol". For conditioning on an event B, he asserts that "we are assuming B is in your probability model, i.e., in the field of events to which you assign probabilities.

So it is implicit in the principle of total evidence that your probability model should include a model for what you learn". On the other hand, Hutchison [12,13] emphasizes that the updating process needs to take into account the circumstances under which the truth of I was conveyed. Also, Bar–Hillel and Falk [14] claims that knowing how knowledge was obtained is "a crucial ingredient to select the appropriate model". These authors present different views about the concept of conditionalization, but all agree on the fact that there would not be a problem if it was known how the information I had become available and therefore, one could build a model including I.

The concept of conditional probability distributions is certainly appropriate as a procedure to update probabilities on the basis of any new information that has already been included in the probability model. However, it can be difficult to construct a model that considers all possible relevant information that could become available in the future. Therefore, a problem arises when one obtains some new and possibly unexpected information and wants to use it to update a probability distribution. Indeed, it does not seem appropriate to assess the probability of something which has been already observed. Our basic assumption is that the information I can be connected to the outcome of interest via a loss function $l(\omega, I)$. In this way, it is possible to update the probability P, even if I is some new unexpected information that was not included in the probabilistic framework.

Clearly, we are dealing with information about outcomes of interest which take many forms. It is well-known that Bayesian inference can rely on expert opinions which often take the form of non-stochastic information. In particular, our framework allows coherent updates of probabilities involving such information, and hence, the practical relevance of our framework to Bayesian inference.

2. Results

This section reports the current definition of conditional probability and presents and motivates our definition for conditional probability with non-stochastic information.

2.1. The New Definition

If $p(\omega)$ represents prior beliefs about ω, then we argue that a valid and coherent update of $p(\cdot)$ on the basis of I is the posterior, $p_I(\cdot)$, where

$$p_I(\omega) = \frac{\exp\{-\lambda l(\omega, I)\}\, p(\omega)}{\sum_{\omega' \in \Omega} \exp\{-\lambda l(\omega', I)\}\, p(\omega')}, \tag{3}$$

and λ is a positive constant.

Looking at this form, it seems that we are proposing the equivalent of a Bayes rule with

$$p(I|\omega) = \exp\{-\lambda l(\omega, I)\}. \tag{4}$$

This is correct only if the piece of information I is an element of a known space of possible outcomes. Indeed, only in such a case would it be possible to assess a probability distribution in such a space.

In our proposal, a much less complex framework is considered, since it is required to just set the loss $l(\omega, I)$ once I has been received. Our axioms which lead to (3) do not regard (4) as the probability.

We shall now go into the details of how some natural assumptions imply (3). The following axioms are considered:

1. The posterior form is given by
$$p_I(\cdot) = \psi(l(\cdot, I), p(\cdot)),$$
for some function (ψ). Since $l(\cdot, I)$ and $p(\cdot)$ are all that are available and set, it is clear that the update must solely depend on these functions. We are, hence, asking for the unique ψ which provides the update for all Ω, i.e., Ω invariant. This is a reasonable requirement since how one updates should not depend on Ω.

2. If the piece of information (I) is made of two independent pieces of information, I_1 and I_2, in the sense that $l(\omega, I) = l(\omega, I_1) + l(\omega, I_2)$, for every ω in Ω, then the posterior probability satisfies the following:
$$\psi\Big(l(\cdot, I_2), \psi\big(l(\cdot, I_1), p(\cdot)\big)\Big) \equiv \psi\big(l(\cdot, I_1) + l(\cdot, I_2), p(\cdot)\big). \tag{5}$$
This ensures we end up with $p_{I_1,I_2}(\omega)$ as the same object whether we update with (I_1, I_2) together or $\{I_1, I_2\}$ one after the other.
3. If $l(\omega, I) = \infty$ for every $\omega \in B^c$ and some $B \subset \Omega$, then
$$\psi(l(\cdot, I), p(\cdot))(\omega) = \begin{cases} \psi(l(\cdot, I)|_B, p_{J(B)}(\cdot))(\omega) & \text{if } \omega \in B \\ 0 & \text{if } \omega \notin B, \end{cases} \tag{6}$$
where $l(\cdot, I)|_B$ is the restriction of $l(\cdot, I)$ to B, $J(B)$ is the information that B certainly occurs or has occured, and $p_{J(B)}$ is p restricted and normalized to B, i.e., $p_{J(B)}(\omega) = p(\omega)\mathbf{1}(\omega \in B)/\sum_{\omega \in B} p(\omega)$. In other words, outcomes with infinite loss are disregarded in the updating.
4. Lower evidence (larger loss) for a state should yield smaller posterior probabilities under the same prior. So, if for some $A \subset \Omega$, $l(\omega, I_1) > l(\omega, I_2)$ for $\omega \in A \subset \Omega$ and $l(\omega, I_1) = l(\omega, I_2) < \infty$ for $\omega \in A^c$, then
$$\textstyle\sum_{\omega \in A} \psi(l(\cdot, I_1), p(\cdot))(\omega) < \sum_{\omega \in A} \psi(l(\cdot, I_2), p(\cdot))(\omega).$$
5. If $l(\omega, I) \equiv$ constant, then $\psi(l(\cdot, I), p(\cdot)) \equiv p(\cdot)$. That is, if the observation provides no information about ω, since the loss function is a constant, then the posterior is the same as the prior.

As a consequence of these, we have the following theorem:

Theorem 1. *For $|\Omega| < \infty$, with axioms 1 to 5, it is uniquely implied that*
$$p_I(\omega) = \frac{\exp\{-\lambda l(\omega, I)\}\, p(\omega)}{\sum_{\omega' \in \Omega} \exp\{-\lambda l(\omega', I)\}\, p(\omega')}, \tag{7}$$
for some $\lambda > 0$.

In order to prove Theorem 1, the following Lemma is useful.

Lemma 1. *Let $|\Omega| < \infty$. Axioms 1 to 5 imply:*

6. *If $\widetilde{l}(\cdot, I) \equiv l(\cdot, I) + c$ for some constant c, then $\psi(\widetilde{l}(\cdot, I), p(\cdot)) \equiv \psi(l(\cdot, I), p(\cdot))$.*
7. *for any set $B \subset \Omega$,*
$$\frac{\psi(l(\cdot, I), p(\cdot))(\omega)}{\sum_{\omega' \in B} \psi(l(\cdot, I), p(\cdot))(\omega')} = \psi(l(\cdot, I)|_B, p_{J(B)}(\cdot))(\omega), \tag{8}$$
for every ω in B. This means that whether we update the prior probability $p(\cdot)$ restricted to set B, or update the prior probability $p(\cdot)$ and then restrict to set B, we obtain the same update.

Proof of Lemma 1. Combination of axioms 2 and 5 yield axiom 6. Indeed, one just needs to consider (5) the case where the loss function $l(\cdot, I_1)$ or $l(\cdot, I_2)$ is constant.

Combining conditions 3 and 5, one obtains that updating with the loss (1) corresponding to information $J(B)$ yields the conditional probability $p_{J(B)}(\omega) = p(\omega)\mathbb{I}_B(\omega)/P(B)$ that is obtained by restricting and normalizing the prior to B. Now, consider $I_1 = J(B)$ and let I_2 be any piece of information associated with a given loss function (l). The sum of the two corresponding losses is infinite on B^c and coincides with l on B, which is the loss considered in condition 3. Therefore, based on condition 2, condition 7 is obtained. □

Proof of Theorem 1. In the two points case, say $\Omega = \{\omega_0, \omega_1\}$, the prior probability ($p(\cdot)$) is determined by a real number (z) in the unit interval (i.e., $z = p(\omega_0)$), and the loss $l(\omega, I)$ takes only two values, say l_0 and l_1 (i.e., $l(\omega_0, I) = l_0, l(\omega_1, I) = l_1$). According to condition 6, we can replace the pair (l_1, l_0) with $(l_1 - l_0, 0)$. Therefore, the posterior probability $p_I(\cdot)$ is a function of $l = l_1 - l_0$ and z, say $\bar{\psi}(l, z)$. In order to proceed, it is convenient to think in terms of odds rather than probabilities. So, if $z = p(\omega_0)$, we can consider $t = z/(1 - z)$ (i.e., $z = t/(1 + t)$). We can do the same with the posterior ($p_I(\omega_0)$). The posterior odds, $p_I(\omega_0)/\{1 - p_I(\omega_0)\}$, are a function of the loss difference (l) and the prior odds (t), which we denote by $\phi(l, t)$.

To deal with the odds, (5) becomes

$$\phi(l + h, t) = \phi(l, \phi(h, t)), \tag{9}$$

where h is a replication of l with a possible different I. Moreover, with a constant loss, the posterior is equal to the prior (condition 5), i.e.,

$$\phi(0, t) = t, \tag{10}$$

for every $t > 0$.

At this stage, we consider a prior with three mass points, say $\Omega = \{\omega_1, \omega_2, \omega_3\}$. The prior is given by $\{z_1, z_2, 1 - z_1 - z_2\}$ (i.e., $p(\omega_1) = z_1$ and $p(\omega_2) = z_2$), with $l(\omega_i, I) = l_i$, for $i = 1, 2, 3$. The loss can be zero at one point without loss of generality (condition 6), so it takes values into the set $\{l_1, l_2, 0\}$. Let us consider the updating rule, ϕ, for priors with just two mass points. We can use this to update the conditional probability of $\{\omega_1\}$ given $\{\omega_1, \omega_3\}$, i.e., $z_1/(z_1 + z_3)$. To this aim, we update the prior with masses $z_1/(z_1 + z_3)$ and $z_3/(z_1 + z_3)$ considering just the loss values, $(l_1, 0)$, i.e., disregarding the point ω_2 with its loss (l_2). In other words, we aim to obtain the right-hand-side of (8) and apply condition 7 given by Lemma 1. $t_{1,3}$ denotes the odds corresponding to the conditional probability of ω_1 given $\{\omega_1, \omega_3\}$, that is $t_{1,3} = z_1/z_3$. $t_{1,3}$ is updated on the basis of the loss values (l_1) and zero, i.e., with $\phi(l_1, t_{1,3})$. Similarly, we define $t_{1,2} = z_1/z_2$ and $t_{2,3} = z_2/z_3$. We update $t_{1,2}$ on the basis of l_1 and l_2, i.e., with $\phi(l_1 - l_2, t_{1,2})$ and $t_{2,3}$ with $\phi(l_2, t_{2,3})$. Clearly, $t_{1,3} = t_{1,2}t_{2,3}$, and this factorization of conditional odds has to hold also after updating, i.e.,

$$\phi(l_1, t_{1,3}) = \phi(l_1 - l_2, t_{1,2})\phi(l_2, t_{2,3}), \tag{11}$$

where $t_{1,3} = t_{1,2} \cdot t_{2,3}$. Formally, this identity is a consequence of (8), i.e., updating the conditional probability is the same as conditioning the updated probability. Since (11) must hold for every $t_{2,3}, t_{1,2} > 0$, and for every $l_1, l_2 \in \mathbb{R}$,

$$\phi(l_1, ts) = \phi(l_1 - l_2, t)\phi(l_2, s),$$

for every $t, s > 0$ and $l_1, l_2 \in \mathbb{R}$. If $l_2 = 0$, and say $l_1 = l$, recalling (10),

$$\phi(l, ts) = \phi(l, t)s,$$

for every $t, s > 0$ and every real l. Letting $t = 1$, we find that

$$\phi(l, s) = \phi(l, 1)s, \tag{12}$$

for every $s > 0$ and every $l \in \mathbb{R}$. A combination of (9) and (12) yields $\phi(l + h, 1) = \phi(l, \phi(h, 1)) = \phi(l, 1)\phi(h, 1)$ for every $h, l \in \mathbb{R}$ which, in turn, implies by monotonicity that $\phi(l, 1) = \exp(-\lambda l)$ for some $\lambda \in \mathbb{R}$. Indeed, according to condition 4, $\phi(l, 1)$ is a monotone, not an increasing function of l. This implies that λ must be positive. Hence,

$$\phi(l, t) = \exp(-\lambda l)t, \tag{13}$$

for every $l \in \mathbb{R}$ and every $t > 0$. In this way, we are basically done with the two points case.

Let us now consider the general finite case. We want to update the prior $p(\cdot)$ with mass points at $\{\omega_1, \ldots, \omega_m\}$ given by $\{z_1, \ldots, z_{m-1}, 1 - (z_1 + \ldots, z_{m-1})\}$, where $z_1, \ldots, z_{m-1}$ are non-negative and their sum is less than or equal to one, and it is convenient to set $z_m := 1 - (z_1 + \cdots + z_{m-1})$. In terms of odds, the prior is given by the vector $(t_1, \ldots, t_{m-1})$ (being $t_i = z_i/(1 - z_i)$, or equivalently, $z_i = t_i/(1 + t_i) = 1 - 1/(1 + t_i)$, $i = 1, \ldots, m - 1$), where $t_i \geq 0$ and $t_1/(1 + t_1) + \cdots + t_{m-1}/(1 + t_{m-1}) \leq 1$. Here, it is convenient to set $t_m := \{\{1 - t_1/(1 + t_1) + \ldots + t_{m-1}/(1 + t_{m-1})\}^{-1} - 1\}^{-1}$. Moreover, we consider an $\mathbb{R}^{m-1}$ valued function $\boldsymbol{\phi}(\boldsymbol{l}, \boldsymbol{t}) = (\phi_1(\boldsymbol{l}, \boldsymbol{t}), \ldots, \phi_{m-1}(\boldsymbol{l}, \boldsymbol{t}))$, which provides the vector of the updated odds as being $\boldsymbol{l} = (l_1, \ldots, l_m)$ and $\boldsymbol{t} = (t_1, \ldots, t_{m-1})$. Now the question is how we could recover $\boldsymbol{\phi}$ from ϕ, where the latter gives the updating rule for the two points case. Recall the notation used for the conditional odds, i.e., $t_{i,j} = z_i/z_j$, is the odds corresponding to the conditional probability of $\{\omega_i\}$ given $\{\omega_i, \omega_j\}$ for a distinct $i, j = 1, \ldots, m$. We can see that $t_j = z_j / \sum_{i \neq j} z_i = (\sum_{i \neq j} t_{i,j})^{-1}$. According to (8), this identity will also have to be satisfied by the updated odds. Since we update $t_{i,j}$ with $\phi(l_i - l_j, t_{i,j})$, we must have

$$\phi_j(l_1, \ldots, l_m; t_1, \ldots, t_{m-1}) = (\textstyle\sum_{i \neq j} \phi(l_i - l_j, t_{i,j}))^{-1},$$

which by (13) becomes:

$$\phi_j(l_1, \ldots, l_m; t_1, \ldots, t_{m-1}) = \exp(-\lambda l_j)(\textstyle\sum_{i \neq j} \exp(-w l_i) t_{i,j})^{-1}. \tag{14}$$

$t_{i,j} = z_i/z_j$, (14) becomes

$$\phi_j(l_1, \ldots, l_m; t_1, \ldots, t_{m-1}) = \frac{\exp(-\lambda l_j) z_j}{\sum_{i \neq j} \exp(-\lambda l_i) z_i},$$

and the updated probability of $\{\omega_j\}$ is

$$1 - (1 + \phi_j(l_1, \ldots, l_m; t_1, \ldots, t_{m-1}))^{-1} = \frac{\exp(-\lambda l_j) z_j}{\sum_{i=1}^{m} \exp(-\lambda l_i) z_i}.$$

In this way, we have shown how to extend our coherent updating rule from the two points case to the general finite case. □

The proof implies the existence of $\lambda > 0$, and this is no surprise since loss functions are only defined up to scalar factors. Hence, the setting of λ is a calibration issue and to set this, we need more information that is not associated with any specific information (I). So, $l(\omega)$ is set to be the loss function that contains initial beliefs at the outset before any further information (I) is obtained, and this $l(\omega)$ is defined to be on the same scale as any $l(\omega, I)$. We regard $l(\omega) = l(\omega, I_0)$, where I_0 is the prior information used to set p.

The prior expected loss is then defined by

$$\mathrm{L} = \lambda \sum_{\omega \in \Omega} l(\omega)\, p(\omega).$$

Now, entropy is also regarded as an expected loss based on the self-information loss function $-\log p(\omega)$. Hence, we also have expected loss/entropy which is measured as

$$\mathrm{E} = -\sum_{\omega \in \Omega} p(\omega)\, \log p(\omega),$$

and so calibration is achieved by matching these expected losses, so that

$$\lambda = -\frac{\sum_{\omega\in\Omega} p(\omega)\log p(\omega)}{\sum_{\omega\in\Omega} l(\omega)p(\omega)}.$$

Finally, in this section, we note that it is quite straightfoward to extend the uniqueness argument to all countably infinite Ω, which replaces the uniqueness argument for all Ω. However, we need more work to extend separate uniqueness to the general Ω.

2.2. An Illustration

The loss function l is chosen by the decision-maker on the basis of the available information. Such information sometimes happens to be stochastic, i.e., belonging to a set B of outcomes to which a probability is assigned. If this is the case, one should update the probability by means of the usual conditional probability. It is tantamount to use the loss function (1) in (3). If the available information is not stochastic, then one can resort to the approach described in the present paper to properly assess the loss function l. A simple and very concrete example is now presented. Consider a horse race in which six horses participate. In order to decide how to bet, one assesses the probability for each horse to win. $p(j)$ demotes the probability that horse number j wins for $j \in \{1,\ldots,6\}$. In this example, the elements of Ω are the numbers corresponding to the horses participating in the race, of which one will win.

Before the race begins, it starts raining. Since conditions have changed, the probabilities need to be updated. It is problematic to pursue this aim by resorting to the current definition of conditional probability. In fact, this requires knowledge of the probability that it rains and that the horse number j wins. As an alternative, one could calculate the conditional probabilities of a win for each horse by applying Bayes' theorem, which requires the probability that it rains given the win of horse j. However, it is raining and the race has not yet been run!

It is therefore appropriate to resort to the definition of a conditional probability distribution given in this paper. So, $p(j)$ is the prior probabilty that horse j wins and $I =$ is the probability that "it is raining". $l(j)$ be the loss function, i.e., the loss incurred to the bookmaker, that is, the governor of the probabilities, if horse j wins. To elaborate, such an $l(j)$ could be the assessed loss if the bookmaker puts all the takings received or a fixed amount on horse j to win.

The bookmaker then, on the same scale, assesses $l(j, I)$ for each j which adjusts the loss $l(j)$. Obviously, a higher score will be given to those horses whose ability to run is more affected by the rain. In this way, one can use the ideas for setting λ to get $\widetilde{l}(j) = \lambda\, l(j, I)$. The updated probability that the j-th horse wins is revised to

$$p_I(j) = \frac{\exp\{-\widetilde{l}(j)\, p(j)\}}{\sum_{i=1}^{6}\exp\{-\widetilde{l}(i)\, p(i)\}},$$

for $j = 1,\ldots,6$.

3. Discussion

We have established a framework in which we can update probabilities in the light of general, i.e., non-stochastic, information. Given that we cannot connect the information and the outcome of interest via a probability model, we do so through a loss function. Minimizing a cumulative loss function involving the information on one side and the probability distribution on the other yields the updated probability distribution. When the information is stochastic, we employ the self information loss function; the solution then reverts to the standard definition of conditional probability.

We believe the framework has direct implications for Bayesian inference where "word of mouth" information from experts is often obtained. It is interesting to note that our framework allows this information to update p at any point. Usually, the non-stochastic information I comes before the

stochastic information given by an event B included in the probability framework, but in our approach, we can have B before I.

Author Contributions: The order of the authors is alphabetical. They contributed equally.

Funding: This research received no external funding.

Acknowledgments: The authors are grateful for the detailed comments of four referees on the revision of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bissiri, P.G.; Walker, S.G. Converting information into probability measures with the Kullback–Leibler divergence. *Ann. Inst. Stat. Math.* **2012**, *64*, 1139–1160. [CrossRef]
2. Bissiri, P.G.; Holmes, C.C.; Walker, S.G. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2016**, *78*, 1103–1130. [CrossRef] [PubMed]
3. Van Fraassen, B.C. The geometry of opinion: Jeffrey shifts and linear operators. *Philos. Sci.* **1992**, *59*, 163–175. [CrossRef]
4. Shafer, G. Jeffrey's rule of conditioning. *Philos. Sci.* **1981**, *48*, 337–363. [CrossRef]
5. Skyrms, B. Maximum entropy inference as a special case of conditionalization. *Synthese* **1985**, *63*, 55–74. [CrossRef]
6. Domotor, Z. Probability kinematics, conditionals, and entropy principles. *Synthese* **1985**, *63*, 75–114. [CrossRef]
7. Diaconis, P.; Zabell, S. Updating Subjective Probability. *J. Am. Stat. Assoc.* **1982**, *77*, 822–830. [CrossRef]
8. Shore, J.; Johnson, R. Axiomatic derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Trans. Inf. Theory* **1980**, *IT-26*, 26–37. [CrossRef]
9. Freund, J.E. Puzzle or paradox? *Am. Stat.* **1965**, *19*, 29–44.
10. Gardner, M. *The Scientific American Book of Mathematical Puzzles and Diversions*; Simon and Schuster: New York, NY, USA, 1959.
11. Shafer, G. Conditional probability. *Int. Stat. Rev.* **1985**, *53*, 261–275. [CrossRef]
12. Hutchison, K. What are conditional probabilities conditional upon? *Br. J. Philos. Sci.* **1999**, *50*, 665–695. [CrossRef]
13. Hutchison, K. Resolving some puzzles of conditional probability. *Adv. Sci. Lett.* **2008**, *1*, 212–221. [CrossRef]
14. Bar-Hillel, M.; Falk, R. Some teasers concerning conditional probabilities. *Cognition* **1982**, *11*, 109–122. [CrossRef]