A definition of conditional probability distribution
with non–stochastic information

Pier Giovanni Bissiri * and Stephen G. Walker †

January 12, 2013

Abstract

The current definition of a conditional probability distribution enables one to update probabilities only on the basis of stochastic information. This paper provides a definition for conditional probability distributions with non–stochastic information. The definition is derived as a solution of a decision theoretic problem, where the information is connected to the outcome of interest via a loss function. We shall show that the Kullback–Leibler divergence plays a central role. Some illustrations are presented.

Keywords: Conditional probability distribution, conditional probability density, loss function, Kullback–Leibler divergence, g-divergence.

2010 Mathematics Subject Classification: 03B48, 60A99; secondary: 62C99.

1 Introduction

The theory of conditional probability distributions is a well-established mathematical theory that provides a procedure to update probabilities taking into account new information. To motivate the new work in this paper, we mention 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 into the probability model.

*Pier Giovanni Bissiri is Postdoctoral researcher, Dipartimento di Statistica, Università degli Studi di Milano–Bicocca, Edificio U7, Via Bicocca degli Arcimboldi 8, Milano 20126, Italy (e-mail: pier.bissiri@unimib.it).

†Stephen G. Walker is Professor, SMSAS, University of Kent, Canterbury, Kent, CT2 7NZ, UK (e-mail: S.G.Walker@kent.ac.uk).
1.1 Notation

Before proceeding, we introduce the notation. Let $Y$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which will be the outcome of interest, and valued into a measurable space $(\mathcal{Y}, \mathcal{Y})$ with probability distribution $P$. Hence, $P$ represents initial belief about the outcome concerning $Y$. By $I$ we shall denote the information obtained about $Y$. If $I$ is stochastic, then we shall represent it by a random variable $X$ from $(\Omega, \mathcal{F}, \mathbb{P})$ into $(\mathcal{X}, \mathcal{X})$ with probability distribution $Q$ and $I$ will be assumed to be an outcome of $X$. We will denote by $\mathbb{P}_I$ the updated $\mathbb{P}$ given information $I$.

We will let $D$ denote the Kullback-Leibler divergence (relative entropy), i.e.,

$$D(Q_1, Q_2) = \int \log \left( \frac{dQ_1}{dQ_2} \right) dQ_1,$$

for any couple $(Q_1, Q_2)$ of probability measures such that $Q_1 \ll Q_2$. More generally we define the $g$-divergence:

$$D_g(Q_1, Q_2) = \int g \left( \frac{dQ_1}{dQ_2} \right) dQ_2,$$

for any couple $(Q_1, Q_2)$ of probability measures such that $Q_1 \ll Q_2$, where $g$ is a convex function from $(0, \infty)$ into $\mathbb{R}$ such that $g(1) = 0$. This class of probability discrepancies has been introduced and studied independently by Ali & Silvey (1966) and Csiszár (1967). The Kullback–Leibler divergence is a particular case, which can be obtained taking $g(x) = x \log(x)$.

1.2 Mathematical framework

When the standard definition of conditional probability does not apply, for reasons we discuss later, we present an alternative definition based on a mathematical decision theoretic framework. 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 outcome of interest $Y$. We do so using loss functions.

The purpose of this paper is to provide a definition of a conditional distribution of $Y$ on the basis of $I$, which we shall denote by $P_I$. We take the pair $(I, P)$ to $P_I$ as the solution to a decision problem based on the minimization of a cumulative loss function. This loss function will be defined on the class of probability measures on $\mathcal{Y}$ that are absolutely continuous with respect to $P$, call this $\mathcal{P}$. Indeed, the conditional probability should be zero on every event whose unconditional probability is zero. Here, $\lambda \in \mathcal{P}$.
will denote the action and the best choice, i.e. minimizing the loss function, will be defined as the conditional probability distribution for $Y$ given $I$. In order to properly assess the loss function, it will be expressed as the sum (cumulative loss) of two terms, i.e.

$$L(\lambda) = H_I(\lambda) + l(\lambda, P),$$

where $l(\lambda, P)$ is a discrepancy between the probability measure $\lambda$ and $P$ and $H_I(\lambda)$ is the component of the loss that takes into account the information relating to $I$. In fact, we will show that $l(\lambda, P)$ should be the Kullback–Leibler divergence for coherence purposes. So, $P_I$ will be defined as that $\lambda$ which minimizes $L(\lambda)$.

### 1.3 Relation to the literature

In the literature, 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 the information can be given in the form of a constraint (or a combination of constraints) on the probability. Constraints considered are of the type

$$\int_{\mathcal{Y}} g(y)\lambda(dy) > 0,$$

where $g$ is a measurable real function on $\mathcal{Y}$ and the strict inequality is sometimes replaced by a not strict one. The idea is to minimize $D(\lambda, P)$ subject to the constraint (2), which represents information $I$. This problem can be solved, i.e. $P_I$ can be obtained, by minimizing $D(\lambda, P)$ subject to the constraint using Lagrange multipliers.

Such a procedure of condizionalization is a specific case in our approach. In fact, it is equivalent to minimize the loss function (1) taking $l$ equal to the Kullback–Leibler divergence and

$$H_I(\lambda) = \begin{cases} 0 & \text{if } \int_{\mathcal{Y}} g(y)\lambda(dy) > 0, \\ +\infty & \text{if } \int_{\mathcal{Y}} g(y)\lambda(dy) \leq 0. \end{cases}$$

For more details about conditionalization based upon constraints on the conditional distribution, see Van Fraassen (1992), Skyrms (1985), Domotor (1985), Diaconis & Zabell (1982) and Shore & Johnson (1980). Our approach is different as we encompass potentially arbitrary information about $Y$, so as long as it is possible to construct a loss function $h_I(y)$ for each $Y$ given $I$. 

3
1.4 Motivation

The random variable $Y$ represents an unknown quantity to which a probability distribution has been assigned and needs to be updated on the basis of new information $I$. If $I$ coincides with an outcome of another random variable $X$, then it is possible to update the unconditional distribution of $Y$ to the probability distribution of $Y$ given $X$. However, to do this, it is required to know all the possible alternatives of $I$, that is, all the outcomes of $X$. Moreover, it is required to assess the joint distribution of $X$ and $Y$ or the conditional distribution of $X$ given $Y$. This is quite easy if, for instance, $I$ is known to be an outcome of some well-defined random experiment. In many situations, one has seen the outcome $X$ and in order to establish an update of the distribution of $Y$, one needs to retrospectively ponder and imagine a joint probability model.

This difficulty arises in different puzzles such as, for instance, Freund’s puzzle of the two aces, introduced by Freund (1965). For other puzzles about conditional probabilities, see, for instance, Gardner (1959).

These puzzles have been widely used to discuss the concept of conditional probability. Hutchison (1999, 2008) emphasizes that the updating process needs to take into account the circumstances under which the truth of $I$ was conveyed. Also, Bar-Hillel & Falk (1982) claim that to know how the knowledge was obtained is “a crucial ingredient to select the appropriate model”. These scholars 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 $I$ became 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 was already included in the probability model. But it can be difficult to construct a model that considers all possible relevant information that in the future could become available. Therefore, the 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 $H_I$ defined on the set of all possible outcomes of $Y$. The conditional distribution of $Y$ given $I$ will be defined as the one that minimizes a cumulative loss in the form given by (1). In this way, it is possible to update the distribution of $Y$, even if $I$ is some new unexpected information, which was not included in the probabilistic
framework. It will be shown that if instead $I$ is the outcome of a random variable $X$ and there is a joint density $f$ for $(X, Y)$, then one can recover as particular case the conditional distribution of $Y$ given $X$. To do this, $l$ is taken to be the Kullback–Leibler divergence. It will be proved that in general it is necessary for the updating procedure to be coherent that $l$ is the Kullback–Leibler divergence.

1.5 Description of the paper

Section 2 contains the main results. In Section 3 some examples will be considered. One such is as follows: assume that $Y$ is a scalar quantity and one learns that $Y$ is close to zero. An answer will be given to this question: how could one update the distribution of $Y$ after learning such information? Section 4 contains a discussion.

2 Defining conditional probability distributions with non–stochastic information

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

2.1 The current definition

In probability theory, a conditional distribution of $Y$ given $X$ is a map $p$ from $\mathcal{Y} \times \mathcal{X}$ into $\mathbb{R}$ such that:

- for each $x$ in $\mathcal{X}$, $p(\cdot, x)$ is a probability measure on $\mathcal{Y}$,
- for each $B$ in $\mathcal{Y}$, $p(B, X(\omega))$ is a version of the conditional probability $\mathbb{P}(Y \in B \mid X(\omega))$, i.e. for each $A$ in $\mathcal{X}$ and each $B$ in $\mathcal{Y}$,

$$\mathbb{P}\{X \in A, Y \in B\} = \int_A p(B, x) \, dQ(x), \quad (3)$$

where $Q$ denotes the probability distribution of $X$.

The conditional distribution is known to be essentially unique, i.e. unique only up to a.s. equality. This is a consequence of $X$ being stochastic. In fact, as Feller (1971, page 160) points out, if, for instance, the distribution of $X$ is concentrated on a subset $\mathcal{X}_0$ of $\mathcal{X}$, no natural definition of $p(B, x)$
is possible for \( x \) outside \( X_0 \). Nevertheless, in individual cases, there usually exists a natural choice dictated by regularity requirements.

Moreover, it is well known that conditional distributions do not always exist unless some conditions are satisfied by the spaces \((X, \mathcal{X})\) and \((Y, \mathcal{Y})\). For more information about conditional probability distributions, see, for instance, Feller (1971) or Billingsley (1995).

This paper will consider the case in which there are two \( \sigma \)-finite measures \( \mu \) and \( \nu \) on \( \mathcal{F} \) such that the probability distribution of \((X, Y)\) is absolutely continuous with respect to \( \mu \times \nu \). Denote its density \( f \). This is a general framework which includes most applications and enables to find easily an expression for the conditional distributions. Generally, \( X \) and \( Y \) are subsets of \( \mathbb{R}^k \), for some \( k \), and \( \mu \) and \( \nu \) are the corresponding Lebesgue measure.

If \( f \) is the density of the probability distribution of \((X, Y)\) with respect to \( \mu \times \nu \), then one can take

\[
p(B, x) = \frac{\int_B f(x, y) \nu(dy)}{\int_Y f(x, y) \nu(dy)},
\]

for every \( B \) in \( Y \) and every \( x \) in \( X \) such that

\[
0 < f_X(x) := \int_Y f(x, y) \nu(dy) < \infty.
\]

Note that \( p(\cdot, x) \) is absolutely continuous w.r.t. \( \nu \) and its density is

\[
f_{Y|X}(y|x) := f(x, y)/f_X(x),
\]

for every \( x \) in \( X \) satisfying (5). The density (6), which is called the conditional density of \( Y \) given \( X \), is what is used in most application to find an expression for the conditional distribution. Therefore, (4) deserves to be considered as the “practical definition” of conditional probability distribution. Indeed, it is the natural version of the conditional distribution of \( Y \) given \( X \) whenever a joint density \( f \) exists for \( X \) and \( Y \).

### 2.2 The loss function

Given it is not always possible to relate new information \( I \) to \( Y \) through probability models, instead, we will rely on the use of loss functions to “connect” the information \( I \) to \( Y \). We will deal with the theory first, and then present some examples.

Before proceeding, let us recall that \( q(B, \cdot) \) satisfying (3) can be seen as the solution of a minimization problem whenever \( Y \) is in \( L^2(\Omega, \mathcal{F}, P) \), by
resorting to the theory of Hilbert spaces (see, for instance, Jacod & Protter 2003). Clearly, this approach relies on the joint distribution of $X$ and $Y$ and therefore is not available when $X$ is replaced by some non–stochastic information $I$.

So, our aim is to define a conditional probability distribution as a solution of a decision problem with a fully motivated loss function; connecting the action, i.e. the conditional distribution, with current and given pieces of information: namely the probability distribution $P$ of $Y$ and $I$, respectively.

The form of the loss function we consider is (1). In particular, $H_I(\lambda)$ will be taken in the integral form i.e. the average or expected loss

$$H_I(\lambda) = \int_Y h_I(y) \, \lambda(dy),$$

where $h_I(\cdot, P)$ is a loss function defined on $Y$. It is more reasonable to assess the loss relating to $Y$ and therefore it is reasonable to be able to construct $h_I(y)$. Examples will be considered later. If $\lambda$ then represents beliefs about $Y$, it is appropriate to consider the expected loss here. Therefore, to define conditional distributions, a cumulative loss will be used of the following form:

$$\int_Y h_I(y) \, \lambda(dy) + l(\lambda, P). \quad (7)$$

This general cumulative loss then represents or assesses the loss to the decision maker if they select probability measure $\lambda$ in the presence of information $I$ and $P$.

2.3 Stochastic information

Let us see how this works when indeed $I$ is equivalent to a random variable $X$ and there is a joint density $f$ for $(X, Y)$. In this setting, the conditional distribution (4) arises as the solution of a decision theoretic problem. To see this, for every $x$ in $\mathbb{X}$ satisfying (5), define the following loss function $\bar{L}_x$:

$$\bar{L}_x(\lambda) := -\int_S \log(f(x, y)/f_Y(y)) \, \lambda(dy) + D(\lambda, P), \quad (8)$$

where

$$f_Y(y) := \int_{\mathbb{X}} f(x, y) \, \mu(dx),$$

$S$ is the set of all $y$ in $S$ such that $0 < f_Y(y) < \infty$, $P$ is the probability distribution of $Y$, $\lambda$ is a probability measure on $\mathbb{Y}$ absolutely continuous
w.r.t. $P$, and $D$. The loss (8) is of the form (7) with
\[ l(\lambda, P) = D(\lambda, P), \]
and
\[ h_I(y) = h(y, x) = -\mathbb{I}_S(y) \log(f(x, y) / f_Y(y)) = -\mathbb{I}_S(y) \log f_{X|Y}(x|y), \tag{9} \]
where $\mathbb{I}_S(y)$ is equal to 1 or 0 depending on whether $y$ belongs to $S$ or not.

For every $x$ in $X$ satisfying (5), the conditional distribution $p(\cdot, x)$ given by (4) minimizes the loss $\bar{L}_x$, since
\[ \bar{L}_x(\lambda) = D(\lambda, p(\cdot, x)) - \log \left( \int_Y f(x, y) \nu(dy) \right). \]

In the loss (8), the first addendum depends on the joint density function of $X$ and $Y$ and therefore, to be able to define such loss, $X$ needs to be stochastic. In other words, a probability distribution has to be assigned to $X$.

The loss (9) is known as the self–information loss function and the most commonly used when $x$ has come from a specified family of densities. So, $H_I$ turns out to be the the expected or average loss, using the self–information loss function $- \log f_{X|Y}(x|y)$.

### 2.4 Non–stochastic information

If the random variable $X$ is replaced by some non–stochastic information $I$, then the self–information loss (9) cannot be defined, but one can still resort to a loss function of the form (7), assessing $h_I(y)$ in a different way. As usual, $h_I(y)$ evaluates the additional loss in outcome $y$ due to the acquirement of $I$. Some examples for this will be considered later.

In the loss (8), the Kullback–Leibler divergence from the marginal of $Y$ can also be replaced by a more general discrepancy, such as the $g$-divergence.

This leads us to consider a more general loss function than (8) as follows:
\[ \int_Y h_I(y) \lambda(dy) + D_g(\lambda, P), \tag{10} \]
where $h_I$ is assessed after learning $I$, information which does not need to be stochastic. As the loss (8), the loss (10) is defined on the class of probability measures on $Y$ that are absolutely continuous with respect to $P$, which is
reasonable. Assume there is a unique probability measure that minimizes (10) in the class of probability measures on $Y$ absolutely continuous with respect to $P$. Then, it will be called the conditional distribution of $Y$ given the information $I$ (according to the discrepancy $D_g$ and the loss $h_I$) and it will be denoted by $P_I$.

At this stage, assume that another piece $J$ of information is available in addition to $I$ and that $I$ and $J$ are not overlapping pieces of information. This happens, for instance, in the stochastic case when $I$ and $J$ are outcomes of two independent random variables. We shall write $IJ$ (or equivalently $JI$) to denote the information obtained combining $I$ with $J$. Being $I$ and $J$ not overlapping, we choose $h_I$, $h_J$ and $h_{IJ}$ satisfying the following additivity property:

$$h_{IJ}(y) = h_I(y) + h_J(y). \quad (11)$$

Clearly, updating the distribution $P$ on the basis of $I$ and $J$ and updating the conditional distribution $P_I$ on the basis of $J$ only, should yield the same probability distribution for $Y$. In the first case, the updated probability distribution is obtained by minimizing the loss:

$$\int_Y h_{IJ}(y) \lambda(dy) + D_g(\lambda, P). \quad (12)$$

In the second one, the loss to minimize is:

$$\int_Y h_J(y) \lambda(dy) + D_g(\lambda, P_I). \quad (13)$$

The two losses (12) and (13) should yield the same updated probability distribution for $Y$.

For this coherence condition to be in force, it is necessary that the discrepancy $D_g$ is the Kullback-Leibler divergence. To be more precise, the following theorem can be stated:

**Theorem.** Let $\tilde{P} := P_I$, and assume that (11) holds and

$$P_{IJ} = \tilde{P}_J, \quad (14)$$

for every probability measure $P$ on $Y$ and for every choice of the loss functions $h_I$ and $h_J$ such that $P_I$, $P_{IJ}$ and $\tilde{P}_J$ are all properly defined.

Then $D_g$ is the Kullback-Leibler divergence.

**Proof.** This result is proven from a different starting point in Bissiri & Walker (2011, Theorem 2.5). Here, a shorter proof is given by assuming the differentiability of $g$. 

9
Assume that \( \mathbb{Y} \) contains at least two distinct points, say \( y_0 \) and \( y_1 \). Otherwise, \( P \) is degenerate and the thesis is trivially satisfied.

To prove this theorem, it is sufficient to consider a very specific choice for \( P \), taking \( P = p_0 \delta_{y_0} + (1 - p_0) \delta_{y_1} \), where \( 0 < p_0 < 1 \). Any probability measure \( \lambda \ll P \) has to be equal to \( p_0 \delta_{y_0} + (1 - p) \delta_{y_1} \), for some \( 0 \leq p \leq 1 \). Therefore, in this specific situation, the loss (10) becomes:

\[
I(p, p_0, h_I) := p h_I(y_0) + (1 - p) h_I(y_1) \\
+ p_0 g \left( \frac{p}{p_0} \right) + (1 - p_0) g \left( \frac{1 - p}{1 - p_0} \right).
\]

Denote by \( p_1 \) the probability \( P_I(\{y_0\}) \), i.e. the minimum point of \( l(p, p_0, h_I) \) as a function of \( p \), and by \( p_2 \) the probability \( P_{I,J}(\{y_0\}) \). By hypotheses, \( p_2 \) is the unique minimum point of both loss functions \( l(p, p_1, h_J) \) and \( l(p, p_0, h_{I,J}) \). Again by hypothesis, we shall consider only those functions \( h_I \) and \( h_J \) such that each one of the functions \( l(p, p_0, h_I) \), \( l(p, p_1, h_J) \), and \( l(p, p_0, h_{I,J}) \), as a function of \( p \), has a unique minimum point, which is \( p_1 \) for the first one and \( p_2 \) for the second and third one. The values \( p_1 \) and \( p_2 \) have to be strictly bigger than zero and strictly smaller than one: this was proved by \cite{Bissiri} Lemma 2). Hence, \( p_1 \) has to be a stationary point of \( l(p, p_0, h_I) \) and \( p_2 \) of both the functions \( l(p, p_1, h_J) \) and \( l(p, p_0, h_{I,J}) \). Therefore,

\[
g' \left( \frac{p_1}{p_0} \right) - g' \left( \frac{1 - p_1}{1 - p_0} \right) = h_I(y_1) - h_I(y_0), \quad (15) \\
g' \left( \frac{p_2}{p_0} \right) - g' \left( \frac{1 - p_2}{1 - p_0} \right) = h_{I,J}(y_1) - h_{I,J}(y_0), \quad (16) \\
g' \left( \frac{p_2}{p_1} \right) - g' \left( \frac{1 - p_2}{1 - p_1} \right) = h_J(y_1) - h_J(y_0). \quad (17)
\]

Recall that \( h_{I,J} = h_J + h_I \) by (11). Therefore, summing up term by term (15) and (17), and considering (16), one obtains:

\[
g' \left( \frac{p_2}{p_0} \right) - g' \left( \frac{1 - p_2}{1 - p_0} \right) \\
= g' \left( \frac{p_1}{p_0} \right) - g' \left( \frac{1 - p_1}{1 - p_0} \right) + g' \left( \frac{p_2}{p_1} \right) - g' \left( \frac{1 - p_2}{1 - p_1} \right). \quad (18)
\]

Recall that by hypothesis (15) \( - (17) \) need to hold for every two functions \( h_I \) and \( h_J \) arbitrarily chosen with the only requirement that \( p_1 \) and \( p_2 \) uniquely exist. Hence, (18) needs to hold for every \( (p_0, p_1, p_2) \) in \((0, 1)^3\). By
substituting $t = p_0$, $x = p_1/p_0$ and $y = p_2/p_1$, (18) becomes

$$g'(xy) - g'\left(\frac{1 - txy}{1 - t}\right) = g'(x) - g'\left(\frac{1 - tx}{1 - t}\right) + g'(y) - g'\left(\frac{1 - txy}{1 - tx}\right),$$

which holds for every $0 < t < 1$, and every $x, y > 0$ such that $x < 1/t$ and $y < 1/(xt)$. Being $g$ convex and differentiable, its derivative $g'$ is continuous. Therefore, letting $t$ go to zero, (19) implies that

$$g'(xy) = g'(x) + g'(y) - g'(1)$$

holds true for every $x, y > 0$. Define the function $\varphi(\cdot) = g'(\cdot) - g'(1)$. This function is continuous, being $g'$ such, and by (20), $\varphi(xy) = \varphi(x) + \varphi(y)$ holds for every $x, y > 0$. Hence, $\varphi(\cdot)$ is $k \ln(\cdot)$ for some $k$, and therefore

$$g'(x) = k \ln(x) + g'(1),$$

where $k = (g'(2) - g'(1))/\ln(2)$. Being $g$ convex, $g'$ is not decreasing and therefore $k \geq 0$. If $k = 0$, then $g'$ is constant, which is impossible, otherwise, for any $h_I$, $p_1$ satisfying (15), either would not exist or would not be unique. Therefore, $k$ must be positive. Being $g(1) = 0$ by assumption, (21) implies that $g(x) = k x \ln(x) + (g'(1) - k)(x - 1)$. Hence,

$$D_g(Q_1, Q_2) = k \int \ln \left( \frac{dQ_1}{dQ_2} \right) dQ_1$$

holds true for some $k > 0$ and for every couple of measures $(Q_1, Q_2)$ such that $Q_1 \ll Q_2$.

In virtue of this theorem, the conditional distribution of $Y$ given the information $I$ is coherent only if it minimizes the loss

$$\bar{L}(\lambda) := \int_Y h_I(y) \lambda(dy) + k \int \ln \left( \frac{d\lambda}{d\nu} \right) d\lambda,$$

where $k$ is some positive constant. To define the loss (22), one needs to assess $h_I$ and $k$. Notice that a probability distribution that minimizes $\bar{L}(\lambda)$, or equivalently $\bar{L}(\lambda)/k$, is uniquely identified by $h_I/k$. In other words, assessing $h_I = h_0$ and $k = k_0$ is equivalent to assess $h_I = h_0/k_0$ and $k = 1$. For this reason, from now on, it will be convenient to fix $k = 1$. 

11
In what follows, only coherent conditional distributions will be considered. Therefore, \( D_g \) will always be assessed to be the Kullback–Leibler divergence. Whenever a probability measure that minimizes (22) (with \( k = 1 \)) exists and is unique, it will be called the conditional probability distribution of \( Y \) given \( I \) and will be denoted by \( P_I \).

If

\[
\int_Y e^{-h_I(y)} P(dy) < \infty,
\]

then \( P_I \) is properly defined and is equal to

\[
P_I(A) = \frac{\int_A e^{-h_I(y)} P(dy)}{\int_Y e^{-h_I(u)} P(du)},
\]

for every measurable subset \( A \) of \( Y \). In fact,

\[
\bar{L}(\lambda) = D(\lambda, P_I) - \ln \left( \int_Y e^{-h_I(y)} P(dy) \right)
\]

holds true for every probability measure \( \lambda \) on \( \mathcal{M} \) such that \( \lambda \ll P \).

By (24), it is clear that the choice of the Kullback–Leibler divergence for \( D_g \) and of a loss \( h_I \) satisfying (23) is sufficient for the coherence condition (14). Moreover, notice that \( P_I \) is defined to be a unique probability measure, not just essentially unique.

3 Illustrations

The loss function \( h_I \) is chosen by the decision-maker on the basis of the available information. Such information sometimes happens to be stochastic, i.e., to belong to a set of outcomes to which a probability is assigned. If this is the case, one should update the probability distribution of \( Y \) by means of the usual conditional distribution. Whenever there is a joint density \( f \) for \( X \) and \( Y \), this is tantamount to use the self-information loss function \( h_I(y, x) = -\ln f_{X|Y}(x|y) \). If the available information is not stochastic, then one can resort to the approach described in the present paper, properly assessing the loss function \( h_I \). To see a practical and simple example, consider the situation mentioned in the Introduction:

**Example 1.** \( Y \) is a scalar quantity and the information \( I \) is that \( Y \) is close to zero. In this case, it is natural to assess:

\[
h_I(y) = wy^2,
\]
where \( w \) is some positive constant, and the conditional distribution of \( Y \) given \( I \) is

\[
P_I(A) = \frac{\int e^{-wy^2} P(dy)}{\int e^{-wy^2} P(dy)}.
\]

**Example 2.** While for the second example everyone would know how to deal with, there is currently no formal mathematical mechanism for pursuing a conditional update. So suppose it becomes known that \( Y \) belongs to \( B \), for some set \( B \). Not because of some preliminary random experiment but rather due to it becoming aware to the decision maker that actually \( B \) is the set of possible values that \( Y \) can take. So the information is non-stochastic. The most natural choice is

\[
h_I(y) = \begin{cases} 
0 & y \in B \\
\infty & y \notin B 
\end{cases}
\]

from which it is easy to deduce that the \( \lambda \) minimising \( \int h_I(y) \lambda(dy) + D(\lambda, P) \) is given by

\[
P_I(A) = \int_{A \cap B} P(dy)/P(B).
\]

This example is relevant to a number of so-called paradoxes whereby it becomes apparent to the decision maker that the outcome space is smaller than the support of \( P \) (e.g. Freund’s paradox of the two aces). How this is learnt is crucial. This has been pointed out by Hutchison (1999, 2008). If the information that \( Y \) belongs to \( B \) is based on some preliminary random experiment, for which a probability model is given, then obviously the unconditional distribution of \( Y \) can be updated resorting to the current definition of conditional probability. If not, there is not currently a rigorous justification for the usage of the conditional probability. The present paper provides a formal and broad enough framework to cover this case. Many philosophers of science, that are mentioned in the Introduction, have discovered paradoxes based on such scenarios.

**Example 3.** To conclude, let us consider a simple and very concrete example. 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. Denote by \( p_j \) the probability that the horse number \( j \) wins, for \( j \in \{1, \ldots, 6\} \). In this example, \( Y \) is the number corresponding to the horse that 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 to know the probability that it rains and that the horse number $j$ wins. As an alternative, one could calculate the conditional probabilities of victory for each horse by applying Bayes' theorem, which requires the probability that it rains given the victory of horse $j$. But it is raining and the race is not yet run!

It is therefore appropriate to resort to the definition of a conditional probability distribution given in this paper. To this aim, one can assess a score to evaluate the disadvantage due to the rain for each horse. Denote by $h_j$ the score referred to the horse number $j$. If the ability of the horse $j$ is unaffected by the rain, then $h_j = 0$. If not, $h_j$ is positive. A higher score will be given to those horses whose ability to run is more affected. In this way, one can set

$$h_I(y) = \sum_{j=1}^{6} h_j I_{\{j\}}(y),$$

where $I$ is the information that it’s raining and $I_0$ is the initial information about the horses and the weather. The updated probability that the $j$-th horse wins turns out to be

$$P_I(\{j\}) = \frac{e^{-h_j}p_j}{\sum_{i=1}^{6} e^{-h_i}p_i},$$

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

4 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.

Acknowledgements

This work was partially supported by ESF and Regione Lombardia, Italy (by the grant “Dote Ricercatori”).
References

Ali, S. M. & Silvey, S. D. (1966), ‘A general class of coefficients of divergence of one distribution from another’, *J. Roy. Statist. Soc. Ser. B* **28**(1), 131–142.

Bar-Hillel, M. & Falk, R. (1982), ‘Some teasers concerning conditional probabilities’, *Cognition* **11**, 109–122.

Billingsley, P. (1995), *Probability and measure*, Wiley Series in Probability and Mathematical Statistics, third edn, John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Bissiri, P. G. & Walker, S. G. (2010), Converting information into probability measures with the Kullback–Leibler divergence. submitted.

Csiszár, I. (1967), ‘Information-type measures of difference of probability distributions and indirect observation’, *Studia Sci. Math. Hungar.* **2**, 229–318.

Diaconis, P. & Zabell, S. (1982), ‘Updating subjective probability’, *J. Am. Statist. Assoc.* **77**, 822–830.

Domotor, Z. (1985), ‘Probability kinematics, conditionals, and entropy principles’, *Synthese* **63**, 75–114.

Feller, W. (1971), *An introduction to probability theory and its applications. Vol. II*, Wiley Series in Probability and Mathematical Statistics, second edn, John Wiley & Sons Inc., New York-London-Sydney.

Freund, J. E. (1965), ‘Puzzle or paradox?’, *Am. Stat.* **19**(4), 29–44.

Gardner, M. (1959), *The scientific American book of mathematical puzzles and diversions*, Simon and Schuster, New York.

Hutchison, K. (1999), ‘What are conditional probabilities conditional upon?’, *Brit. J. Phi. Sci.* **50**, 665–695.

Hutchison, K. (2008), ‘Resolving some puzzles of conditional probability’, *Adv. Sci. Lett.* **1**, 212–221.

Jacod, J. & Protter, P. (2003), *Probability essentials*, second edn, Springer–Verlag, Berlin, Heidelberg, New York.
Shore, J. & Johnson, R. (1980), ‘Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy’, *IEEE Trans. Inf. Theory* **IT-26**(1), 26–37.

Skyrms, B. (1985), ‘Maximum entropy inference as a special case of conditionalization’, *Synthese* **63**, 55–74.

Van Fraassen, B. C. (1992), ‘The geometry of opinion: Jeffrey shifts and linear operators’, *Philos. Sci.* **59**, 163–175.