Synthese
DOI 10.1007/s11229-016-1070-8

Conditioning using conditional expectations: the Borel–Kolmogorov Paradox

Z. Gyenis[1,2] · **G. Hofer-Szabó**[3] · **M. Rédei**[3,4]

Received: 13 March 2015 / Accepted: 7 March 2016

Abstract The Borel–Kolmogorov Paradox is typically taken to highlight a tension between our intuition that certain conditional probabilities with respect to probability zero conditioning events are well defined and the mathematical definition of conditional probability by Bayes' formula, which loses its meaning when the conditioning event has probability zero. We argue in this paper that the theory of conditional expectations is the proper mathematical device to conditionalize and that this theory allows conditionalization with respect to probability zero events. The conditional probabilities on probability zero events in the Borel–Kolmogorov Paradox also can be calculated using conditional expectations. The alleged clash arising from the fact that one obtains different values for the conditional probabilities on probability zero events depending on what conditional expectation one uses to calculate them is resolved by showing that the different conditional probabilities obtained using different conditional expectations cannot be interpreted as calculating in different parametrizations of the conditional probabilities of the same event with respect to the same conditioning conditions. We conclude that there is no clash between the correct intuition about what the conditional

✉ M. Rédei
m.redei@lse.ac.uk

Z. Gyenis
gyz@renyi.hu

G. Hofer-Szabó
szabo.gabor@btk.mta.hu

1 MTA Alfréd Rényi Institute of Mathematics, Budapest, Hungary

2 Department of Algebra, BUTE, Budapest, Hungary

3 Research Center for the Humanities, Budapest, Hungary

4 Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London, UK

probabilities with respect to probability zero events are and the technically proper concept of conditionalization via conditional expectations—the Borel–Kolmogorov Paradox is just a pseudo-paradox.

Keywords Conditionalization · Borel–Kolmogorov Paradox · Interpretation of probability

"The concepts of conditional probability and expected value with respect to a σ-field underlie much of modem probability theory. The difficulty in understanding these ideas has to do not with mathematical detail so much as with probabilistic meaning..." (Billingsley 1995, *p. 427*)

1 The Borel–Kolmogorov Paradox and the main claim of the paper

Suppose we choose a point randomly with respect to the distribution given by the uniform measure on the surface of the unit sphere in three dimension. What is the conditional probability that a randomly chosen point is on an arc of a great circle on the sphere on condition that the point lies on that great circle? Since a great circle has measure zero in the surface measure on the sphere, the Bayes' formula cannot be used to calculate the conditional probability in question. On the other hand one has the *intuition* that the conditional probability of the randomly chosen point lying on an arc is well defined and is proportional to the length of the arc. This tension between the "ratio analysis" (Bayes' formula) of conditional probability and our intuition is known as the Borel-Kolmogorov Paradox. The tension seems to be aggravated by the fact that different attempts to replace the Bayes' formula by other, apparently reasonable, methods to calculate the conditional probability in question lead to different values.

The Borel–Kolmogorov Paradox has been discussed both in mathematical works on probability theory proper (Kolmogorov 1933, pp. 50–51; Billingsley 1995, p. 441; de Finetti 1972, p. 203; Proschan and Presnell 1998; Rao 1988; Rao 2005, p. 65; Seidenfeld et al. 2001), and in the literature on philosophy of probability (Borel 1909, pp. 100–104; Easwaran 2008; Hájek 2003; Jaynes 2003, p. 470; Howson 2014; Myrvold 2014; Rescorla 2014; Seidenfeld 2001). One can discern two main attitudes towards the Borel–Kolmogorov Paradox: a radical and a conservative.

According to radical views, the Borel–Kolmogorov Paradox poses a serious threat for the standard measure theoretic formalism of probability theory, in which conditional probability is a defined concept, and this is regarded as justification for attempts at axiomatizations of probability theory in which the conditional probability is taken as the primitive rather than a defined notion (Hájek 2003; Harper 1975; Fraassen 1976). Such axiomatizations have been given by Popper (1938, 1955, 1995), and Rényi (1955) (see Makinson 2011 for a recent analysis of Rényi's and Popper's approach).

According to "conservative" papers the Borel–Kolmogorov Paradox just makes explicit an insufficiency in naïve conditioning that can be avoided within the measure theoretic framework by formulating the problem of conditioning properly and carefully. Once this is done, the Borel–Kolmogorov Paradox is resolved. Kolmogorov himself took this latter position (Kolmogorov 1933, pp. 50–51). Billingsley (1995, p. 441), Proschan and Presnell (1998, p. 249) and Rao (1988, p. 441) write about

the Borel–Kolmogorov Paradox in the same spirit (Proschan and Presnell call the Borel–Kolmogorov Paradox the "equivalent event fallacy").

The present paper falls into the conservative group: We claim that the Borel–Kolmogorov Paradox is in perfect harmony with measure theoretic probability theory, if one uses conditional expectations as the conditioning device to define conditional probabilities. But we go substantially beyond the treatment of the paradox in the conservative papers in several important respects. We also display what we think are the problematic reasonings and interpretations in the "radical papers", which we see as the main reason why the radical papers take a radical position about the insufficiency of conditionalization in the framework of Kolmogorovian probability theory. The main points in our paper about why and how the paradox disappears naturally from the Borel–Kolmogorov Paradox if one treats it in the spirit of measure theoretic probability theory can be summarized as follows.

Conservative assessments of the status of the Borel–Kolmogorov Paradox (for instance Kolmogorov's resolution Kolmogorov 1933 and Billingsley's short presentation Billingsley 1995) typically just state that one can obtain a conditional probability on a great circle using the theory of conditional expectations if one specifies the conditioning σ-field to be the one defined by (measurable sets of) meridian circles containing the great circle. But this fact, in and by itself, cannot be considered as a complete explanation of how and why the paradox disappears from the Borel–Kolmogorov Paradox, for two reasons. One is that conditional expectations are determined by conditioning σ-fields up to measure zero only. Hence, the conditional probabilities defined by the conditional expectation determined by the σ-field specified by the meridians leave the conditional probability undefined on any *single* great circle—only on great circles forming a non-measure zero set in the surface measure are the conditional distribution determined this way. We will argue however that the product structure of the probability space formed by the sphere with its surface measure together with a special location of the conditioning σ-field with respect to the product structure single out a *particular version* of the conditional expectation that yields conditional distribution on *all* great circles.

The other reason is that it is an essential part of the Borel paradox that the "intuitively correct" conditional probability on the great circle which the Bayes' rule cannot provide is the *uniform* one, and the conditional probability determined by the σ-field defined by the meridian circles is *not* uniform. Papers such as Jaynes (2003), Rescorla (2014), Myrvold (2014) and Howson (2014) do recall a derivation of the *uniform* conditional probability on the great circle, and we will show how one can obtain this "intuitively correct" uniform distribution on great circles by choosing a conditioning σ-field and a version of the corresponding conditional expectation. Being aware of the fact that the non-uniform conditional probability also can and has been derived, the papers Jaynes (2003), Rescorla (2014), Myrvold (2014) and Howson (2014) see the "Description-Relativity Worry" Howson (2014, p. 8) emerge, namely the worry that the conditional probability of events depends on how one describes the random events. We will argue that the Description Relativity Worry is unjustified because it is based on an all-too casual understanding of what the description-(in)dependence of probabilities is. We give a careful analysis of the concept of "re-coordinatization" of random events and of the concept of "re-parametrization" of probability measure spaces, and

we prove that the uniform and non-uniform conditional probabilities obtained using different conditional expectations cannot be interpreted as calculating in different parametrizations of the conditional probabilities of the same events with respect to the same conditioning conditions. A crucial element in this proof is showing explicitly that the σ-fields that determine the conditional expectations yielding different conditional probability distributions on the great circle are non-isomorphic but a properly defined re-coordinatization of a probability space describing a random phenomenon entails isomorphism of the respective σ-fields.

Defusing the Description Relativity Worry does not resolve the tension, however, between the uniform and non-uniform conditional probabilities: it seems that the uniform conditional probability is the intuitively correct conditional probability on a great circle, whereas the non-uniform is not. We will argue however that *both* are (or rather: can be) intuitively correct. The argument is based on specifying concepts and reasonings as probabilistic if they are invariant with respect to isomorphisms of probability measure spaces, and non-probabilistic if they are not invariant. Using this distinction we try to make explicit the reasons why one may have the intuition that the uniform length measure on the arc is *the* correct conditional probability on a great circle. We claim that this intuition is fallacious; although it is typically not questioned in the philosophical literature on the Borel–Kolmogorov Paradox. The error in the intuition is the lack of clean separation of probabilistic and non-probabilistic concepts and reasoning: The intuition that the uniform distribution is the correct one is based on (tacit) symmetry considerations. We will see how these can be made mathematically precise and explicit but we claim they are not invariant with respect to measure theoretic isomorphisms of the probability space occurring in the Borel–Kolmogorov Paradox.

We will conclude that there is nothing paradoxical in the Borel–Kolmogorov Paradox; hence, although one might in principle have good reasons to develop an axiomatization of probability based on the concept of conditional probability as primitive notion, the Borel–Kolmogorov Paradox is not one of them.

The structure of the paper is the following. Section 2 is a concise review of the notion of conditional expectation and the concept of conditional probability defined via conditional expectations. Section 3 describes the conditional expectation in the case when the set of elementary events are the points of the two dimensional unit square with the Lebesgue measure on the square giving the probabilities and when the conditioning Boolean subalgebra is the σ-field generated by the measurable sets of one-dimensional slices of the square. This example is a simplified version of the Borel–Kolmogorov situation without the technical complication resulting from the non-trivial geometry of the sphere; hence the main idea of how one should treat conditional probabilities in the Borel–Kolmogorov situation in terms of conditional expectations can be illustrated on this example with a minimal amount of technicality. Section 4 calculates the "intuitively correct" uniform conditional distribution on a great circle by choosing a particular σ-field in the Borel–Kolmogorov situation. Section 5 calculates the "intuitively problematic" conditional distribution on great circles that are meridian circles with respect to fixed *North* and *South Poles* by using conditional expectations defined by the σ-field determined by measurable sets of these meridian circles. (Details of these calculations are given in the Appendix section.) Section 6 shows that these different conditional distributions do not stand in contradiction; in

particular, it is shown that they cannot, hence should not, be considered as conditional probabilities obtained via different parametrization of the same event with respect to the same conditioning conditions. Section 7 attempts to display the possible roots of the fallacious intuition that only the uniform distribution on great circles is the correct conditional probability. We close the paper by some general comments and specific remarks on Kolmogorov's resolution of the paradox (Sect. 8).

2 Conditional expectation and conditioning

We fix some notation that will be used throughout the paper. $(X, \mathcal{S}, p)$ denotes a probability measure space: X is the set of elementary events, $\mathcal{S}$ is a σ-field of some subsets of X, p is a probability measure on $\mathcal{S}$. The negation of event $A \in \mathcal{S}$ is denoted by $A^{\perp}$.

Given $(X, \mathcal{S}, p)$, the set of p-integrable functions is denoted by $\mathcal{L}^1(X, \mathcal{S}, p)$; elements of this function space are the integrable random variables. The characteristic (indicator) functions χ_A of the sets $A \in \mathcal{S}$ are in $\mathcal{L}^1(X, \mathcal{S}, p)$ for all A. The probability measure p defines a linear functional ϕ_p on $\mathcal{L}^1(X, \mathcal{S}, p)$ given by the integral:

$$\phi_p(f) \doteq \int_X f dp \quad f \in \mathcal{L}^1(X, \mathcal{S}, p) \tag{1}$$

The map $f \mapsto \| f \|_1 \doteq \phi_p(|f|)$ defines a seminorm $\| \cdot \|_1$ on $\mathcal{L}^1(X, \mathcal{S}, p)$ (only a *semi*norm because in the function space $\mathcal{L}^1(X, \mathcal{S}, p)$ functions differing on p-probability zero sets are *not* identified). The linear functional ϕ_p is continuous in the seminorm $\| \cdot \|_1$.

For more details on the above notions (and other mathematical concepts used here without definition) see the standard references for the measure theoretic probability theory (Loéve 1963; Billingsley 1995; Rosenthal 2006; Bogachev 2007). Section 19 in Billingsley (1995) discusses further properties of the function space $\mathcal{L}^1(X, \mathcal{S}, p)$.

2.1 Conditional expectation illustrated on the simplest case

Let $(X, \mathcal{S}, p)$ be a probability space and assume $A \in \mathcal{S}$ is such that $p(A), p(A^{\perp}) \neq 0$. When one conditionalizes with respect to A using the Bayes' rule

$$p(B|A) = \frac{p(B \cap A)}{p(A)} \tag{2}$$

one also (tacitly) conditionalizes on $A^{\perp}$ because the number

$$p(B|A^{\perp}) = \frac{p(B \cap A^{\perp})}{p(A^{\perp})} \tag{3}$$

also is well defined. Thus, one always conditionalizes not just on the single event A but on the four-element Boolean subalgebra $\mathcal{A}$ of $\mathcal{S}$:

$$\mathcal{A} \doteq \{\emptyset, A, A^{\perp}, X\} \tag{4}$$

One can keep track of both of the conditional probabilities (2)–(3) by defining a map T that assigns to the characteristic function χ_B of $B \in \mathcal{S}$ another function $T\chi_B$ defined by

$$T\chi_B \doteq \frac{p(B \cap A)}{p(A)}\chi_A + \frac{p(B \cap A^{\perp})}{p(A^{\perp})}\chi_{A^{\perp}} \tag{5}$$

T takes its value in $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$, where $p_{\mathcal{A}}$ is the restriction of p to $\mathcal{A}$. Since $\mathcal{L}^1(X, \mathcal{S}, p)$ is the closure of the linear combinations of characteristic functions, T can be extended linearly from the characteristic functions of $\mathcal{L}^1(X, \mathcal{S}, p)$ to the whole $\mathcal{L}^1(X, \mathcal{S}, p)$. Denote the extension by $\mathscr{E}(\cdot \mid \mathcal{A})$.

The upshot: The conditionalizations (2)–(3) defined by the Bayes' rule define a linear map

$$\mathscr{E}(\cdot \mid \mathcal{A})\colon \mathcal{L}^1(X, \mathcal{S}, p) \to \mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}}) \tag{6}$$

The function $\mathscr{E}(\cdot \mid \mathcal{A})$ has the following properties:

(i) For all $f \in \mathcal{L}^1(X, \mathcal{S}, p)$, the $\mathscr{E}(f|\mathcal{A})$ is $\mathcal{A}$-measurable.
(ii) $\mathscr{E}(\cdot \mid \mathcal{A})$ preserves the integration:

$$\int_Z \mathscr{E}(f \mid \mathcal{A}) dp_{\mathcal{A}} = \int_Z f dp \quad \forall Z \in \mathcal{A} \tag{7}$$

Definition 1 $\mathscr{E}(\cdot \mid \mathcal{A})$ is called the $\mathcal{A}$-conditional expectation from $\mathcal{L}^1(X, \mathcal{S}, p)$ to $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$.

Note that the $\mathcal{A}$-conditional expectation $\mathscr{E}(\cdot \mid \mathcal{A})$ is a map between function spaces, *not* a probability measure and *not* the expectation value of any random variable. These latter concepts can be easily recovered from T, see below.

2.2 Conditionalization as Bayesian statistical inference illustrated on the simplest case

We argue in this subsection that the proper way of viewing the standard conditionalization (i.e. Bayes' formula) is to interpret it as (a special case of) Bayesian statistical inference, and that to treat general Bayesian statistical inference, conditional expectations are an indispensable concept.

Let $(X, \mathcal{S}, p)$ be a probability space and $A \in \mathcal{S}$ such that $p(A) \neq 0$. The conditional probability $p(B|A)$ given by Bayes' rule defines another probability measure q on $\mathcal{S}$:

$$q(B) \doteq p(B|A) = \frac{p(B \cap A)}{p(A)} \quad \forall B \in \mathcal{S} \tag{8}$$

The conditional probability measure q obviously has the feature that its restriction to the Boolean subalgebra $\mathcal{A} = \{\emptyset, A, A^{\perp}, X\}$ has specific values on A and $A^{\perp}$:

$$q(A) = 1 \tag{9}$$
$$q(A^{\perp}) = 0 \tag{10}$$

Thus the values $q(B)$ of the conditional probability measure q on elements $B \in \mathcal{S},\ B \notin \mathcal{A}$ given by (8) can be viewed as values of the extension to $\mathcal{S}$ of the probability measure on $\mathcal{A}$ that takes on the specific values (9)–(10) on $\mathcal{A}$. To formulate this differently: the definition of conditional probability by Bayes' rule is an answer to the question: If a probability measure is given on $\mathcal{A}$ that has the values (9)–(10), what is its extension from $\mathcal{A}$ to $\mathcal{S}$? This is a particular case of the problem of statistical inference: One can replace the prescribed specific values (9)–(10) by more general ones and ask the same question: Suppose one is given a probability measure $q_{\mathcal{A}}$ on $\mathcal{A}$:

$$q_{\mathcal{A}}(A) = r_A \tag{11}$$
$$q_{\mathcal{A}}(A^{\perp}) = r_{A^{\perp}} = 1 - r_A \tag{12}$$

What is the extension q of $q_{\mathcal{A}}$ from $\mathcal{A}$ to $\mathcal{S}$? Formulated differently: what are the conditional probabilities $q(B)$ of events $B \in \mathcal{S},\ B \notin \mathcal{A}$ on condition that the probabilities $q(A)$ of events $A \in \mathcal{A}$ are fixed and are equal to $q_{\mathcal{A}}(A)$? This is a special case of the problem of statistical inference [see Marchand (1977, 1981) and Marchand (1982) for a detailed discussion of statistical inference and conditionalization].

In general, there is no unique answer to this question, there exist many extensions. *Bayesian statistical inference*, which is based on the standard notion of conditional probability given by Bayes' formula, is one particular answer. This answer presupposes a background probability measure p on $\mathcal{S}$ with respect to which the conditional probabilities $q(B)$ are inferred from $q_{\mathcal{A}}$. To formulate the Bayesian answer properly, one has to re-formulate the question of statistical inference in terms of functional analysis as follows: let $\psi_{\mathcal{A}}$ be the continuous linear functional on $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$ determined by $q_{\mathcal{A}}$ (cf. Equation (1)).

Problem of statistical inference: Given the continuous linear functional $\psi_{\mathcal{A}}$ on $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$, what is the extension of $\psi_{\mathcal{A}}$ from $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$ to a continuous linear functional ψ on $\mathcal{L}^1(X, \mathcal{S}, p)$?

The Bayesian answer:

Definition 2 (Bayesian inference—elementary case). Let the extension ψ be

$$\psi(f) \doteq \psi_{\mathcal{A}}(\mathscr{E}(f \mid \mathcal{A})) \quad \forall f \in \mathcal{L}^1(X, \mathcal{S}, p) \tag{13}$$

where $\mathscr{E}(\cdot \mid \mathcal{A})$ is the $\mathcal{A}$-conditional expectation from $\mathcal{L}^1(X, \mathcal{S}, p)$ to $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$.

Note that $\psi_{\mathcal{A}}$ needs an $\mathcal{A}$-measurable function as input, and $\mathscr{E}(\cdot \mid \mathcal{A})$ is a continuous linear function that takes any $\mathcal{S}$-measurable function to an $\mathcal{A}$-measurable function, so

it makes sense putting them together in (13) to obtain the continuous linear functional that we want. Also note that one has to show/argue that definition (13) does indeed yield an *extension* of $\psi_{\mathcal{A}}$ that is continuous—see the general case in Sect. 2.4.

Remark 1 The above stipulation of Bayesian statistical inference contains the usual Bayesian conditioning of a probability measure: If in (11)–(12) we demand (9)–(10); i.e. that $r_A = 1, r_{A^\perp} = 0$, then for characteristic functions $\chi_B \in \mathcal{L}^1(X, \mathcal{S}, p)$, $B \in \mathcal{S}$, we have:

$$q(B) = \psi(\chi_B) \tag{14}$$

$$= \psi_{\mathcal{A}}(\mathscr{E}(\chi_B \mid \mathcal{A})) = \int_X \mathscr{E}(\chi_B \mid \mathcal{A}) dq_{\mathcal{A}} \tag{15}$$

$$= \int_X \Big[\frac{p(B \cap A)}{p(A)} \chi_A + \frac{p(B \cap A^\perp)}{p(A^\perp)} \chi_{A^\perp} \Big] dq_{\mathcal{A}} \tag{16}$$

$$= \frac{p(B \cap A)}{p(A)} \int_X \chi_A dq_{\mathcal{A}} + \frac{p(B \cap A^\perp)}{p(A^\perp)} \int_X \chi_{A^\perp} dq_{\mathcal{A}} \tag{17}$$

$$= \frac{p(B \cap A)}{p(A)} q_{\mathcal{A}}(A) + \frac{p(B \cap A^\perp)}{p(A^\perp)} q_{\mathcal{A}}(A^\perp) \tag{18}$$

$$= \frac{p(B \cap A)}{p(A)} \underbrace{r_A}_{=1} + \frac{p(B \cap A^\perp)}{p(A^\perp)} \underbrace{r_{A^\perp}}_{=0} \tag{19}$$

$$= \frac{p(B \cap A)}{p(A)} \tag{20}$$

So the Bayesian answer given in terms of the conditional expectation to the general question of statistical inference covers the case when the probability measure $q_{\mathcal{A}}$ defined on the small Boolean subalgebra $\mathcal{A}$ of $\mathcal{S}$ takes on arbitrary values—not just the extreme values $q_{\mathcal{A}}(A) = 1$ and $q_{\mathcal{A}}(A^\perp) = 0$. The notion of conditional expectation is indispensable to cover this general case of Bayesian statistical inference.

2.3 Conditional expectation: the general case

One can generalize the notion of conditional expectation by replacing the four-element Boolean algebra $\mathcal{A}$ generated by a single element A [see Eq. (4)] by an arbitrary sub-σ-field $\mathcal{A}$ of $\mathcal{S}$:

Definition 3 Let $(X, \mathcal{S}, p)$ be a probability space, $\mathcal{A}$ be a sub-σ-field of $\mathcal{S}$, and $p_{\mathcal{A}}$ be the restriction of p to $\mathcal{A}$. A map

$$\mathscr{E}(\cdot \mid \mathcal{A}) \colon \mathcal{L}^1(X, \mathcal{S}, p) \to \mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}}) \tag{21}$$

is called an $\mathcal{A}$-conditional expectation from $\mathcal{L}^1(X, \mathcal{S}, p)$ to $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$ if (i) and (ii) below hold:

(i) For all $f \in \mathcal{L}^1(X, \mathcal{S}, p)$, the $\mathscr{E}(f \mid \mathcal{A})$ is $\mathcal{A}$-measurable.

(ii) $\mathscr{E}(\cdot \mid \mathcal{A})$ preserves the integration on elements of $\mathcal{A}$:

$$\int_Z \mathscr{E}(f \mid \mathcal{A}) dp_{\mathcal{A}} = \int_Z f dp \qquad \forall Z \in \mathcal{A}. \tag{22}$$

It is not obvious that such a map $\mathscr{E}(\cdot \mid \mathcal{A})$ exists but the Radon–Nikodym theorem entails that it *always* does:

Proposition 1 (Billingsley 1995, p. 445; Bogachev 2007 Theorem 10.1.5). *Given any $(X, \mathcal{S}, p)$ and any sub-σ-field $\mathcal{A}$ of $\mathcal{S}$, a conditional expectation $\mathscr{E}(\cdot \mid \mathcal{A})$ from $\mathcal{L}^1(X, \mathcal{S}, p)$ to $\mathcal{L}^1(X, \mathcal{A}, p_{\mathcal{A}})$ exists.*

Note that uniqueness is not part of the claim in Proposition 1, and for good reason: the conditional expectation is only unique up to measure zero:

Proposition 2 (Billingsley 1995, Theorem 16.10 and p. 445; Bogachev 2007, p. 339). *If $\mathscr{E}'(\cdot \mid \mathcal{A})$ is another conditional expectation then for any $f \in \mathcal{L}^1(X, \mathcal{S}, p)$ the two $\mathcal{L}^1$-functions $\mathscr{E}(f \mid \mathcal{A})$ and $\mathscr{E}'(f \mid \mathcal{A})$ are equal up to a p-probability zero set.*

Different conditional expectations equal up to measure zero are called *versions* of the conditional expectation. The claims in the next proposition are to be understood as "up to measure zero".

Proposition 3 (Billingsley 1995, Sect. 34). *A conditional expectation has the following properties:*

(i) *$\mathscr{E}(\cdot \mid \mathcal{A})$ is a linear map.*
(ii) *$\mathscr{E}(\cdot \mid \mathcal{A})$ is a projection:*

$$\mathscr{E}(\mathscr{E}(f \mid \mathcal{A}) \mid \mathcal{A}) = \mathscr{E}(f \mid \mathcal{A}) \qquad \forall f \in \mathcal{L}^1(X, \mathcal{S}, p) \tag{23}$$

Remark 2 If $\mathcal{A}$ is generated by a countably infinite set $\{A_i\}_{i \in \mathbb{N}}$ of pairwise orthogonal elements from $\mathcal{S}$ such that $p(A_i) \neq 0$ $(i = 1, \ldots)$, then the conditional expectation (21) can be given explicitly on the characteristic functions $\mathcal{L}^1(X, \mathcal{S}, p)$ by a formula that is the complete analogue of (5):

$$\mathscr{E}(\chi_B \mid \mathcal{A}) = \sum_i \frac{p(B \cap A_i)}{p(A_i)} \chi_{A_i} \qquad \forall B \in \mathcal{S} \tag{24}$$

However, for a general $\mathcal{A}$ the conditional expectation cannot be given explicitly, its existence is the corollary of the Radon-Nikodym theorem, which is a non-constructive, pure existence theorem. Note also that if $\mathcal{A}$ is generated by a countably infinite set $\{A_i\}_{i \in \mathbb{N}}$ of pairwise orthogonal elements from $\mathcal{S}$ but $p(A_i) = 0$ for some A_i then (24) still yields the conditional expectation with the modification that the undefined $\frac{p(B \cap A_i)}{p(A_i)}$ is replaced by any number—this is the phenomenon of the conditional expectation being defined up to a probability zero set (Proposition 2).

Remark 3 The conditional expectations can be thought of as an averaging or coarse graining process: if the sub-σ-field $\mathcal{A}$ is generated by the disjunct elements A_λ, where $\lambda \in \Lambda$ are parameters in an arbitrary index set (not necessarily countable), in which case A_λ are atoms in the generated σ-field $\mathcal{A}$, then the $\mathcal{A}$-measurability condition on the $\mathcal{A}$-conditional expectation entails that $\mathcal{E}(f \mid \mathcal{A})$ is a constant function on every A_λ. This constant value on A_λ is the averaged, course-grained value of f on $\mathcal{A}_\lambda$. (The event A_λ might very well not be an atom in $\mathcal{S}$, and so f can vary on elements and subsets of A_λ.)

2.4 Bayesian statistical inference and conditional expectation: general case

Problem of statistical inference: general formulation: Let $(X, \mathcal{S}, p)$ be a probability space, $\mathcal{A}$ be a sub-σ-field of $\mathcal{S}$. Assume that $\psi_\mathcal{A}$ is a $||\cdot||_1$-continuous linear functional on $\mathcal{L}^1(X, \mathcal{A}, p_\mathcal{A})$ determined by a probability measure $q_\mathcal{A}$ given on $\mathcal{A}$ via integral [cf. Eq. (1)]. What is the extension ψ of $\psi_\mathcal{A}$ from $\mathcal{L}^1(X, \mathcal{A}, p_\mathcal{A})$ to a $||\cdot||_1$-continuous linear functional on $\mathcal{L}^1(X, \mathcal{S}, p)$?

The Bayesian answer:

Definition 4 (Bayesian statistical inference). Let the extension ψ be

$$\psi(f) \doteq \psi_\mathcal{A}(\mathscr{E}(f \mid \mathcal{A})) \qquad \forall f \in \mathcal{L}^1(X, \mathcal{S}, p) \tag{25}$$

where $\mathscr{E}(\cdot \mid \mathcal{A})$ is the $\mathcal{A}$-conditional expectation from $\mathcal{L}^1(X, \mathcal{S}, p)$ to $\mathcal{L}^1(X, \mathcal{A}, p_\mathcal{A})$.

Note that because $\mathscr{E}(\cdot \mid \mathcal{A})$ is a projection operator on $\mathcal{L}^1(X, \mathcal{S}, p)$ (Proposition 3), ψ is indeed an *extension* of $\psi_\mathcal{A}$, and because $\mathscr{E}(\cdot \mid \mathcal{A})$ is $\|\cdot\|_1$-continuous, the extension ψ also is $||\cdot||_1$-continuous.

The notion of conditional *probability* of an event obtains as a special case of Bayesian statistical inference so defined [see Marchand (1977, 1981) and Marchand (1982) for further discussion of the relation of statistical inference and conditionalization]:

Definition 5 If $B \in \mathcal{S}$ then its $(\mathcal{A}, \psi_\mathcal{A})$-conditional probability $q(B)$ is the expectation value of its characteristic function χ_B computed using the formula (25) containing the $\mathcal{A}$-conditional expectation:

$$q(B) \doteq \psi(\chi_B) = \psi_\mathcal{A}(\mathcal{E}(\chi_B \mid \mathcal{A})) \tag{26}$$

Comments on the definition of conditional probability:

1. Note that there is no restriction in this general definition of conditional probability on the conditioning σ-field $\mathcal{A}$, nor on the values the unconditional (background) measure p can have on this algebra $\mathcal{A}$; in particular some elements of the conditioning σ-field $\mathcal{A}$ can have zero unconditional probability. Thus, in principle, Definition 5 of conditional probability covers such cases and one can have conditional probabilities with respect to events that have prior probability zero.

2. *If* the σ-field $\mathcal{A}$ is generated by a single element A, and *if* element A has non-zero unconditional probability, $p(A) \neq 0$, and *if* the conditional measure is assumed to take value 1 on A, then the conditional probability measure q is the normalized restriction of the unconditional measure p to A; i.e. in this special case the conditional probability is given by the Bayes' rule (see Remark 1). But this special case is not only extremely special but also slightly deceptive because it conceals the true content and conceptual structure of conditionalization: that conditional probabilities depend sensitively on *three* conditions (variables):
   (i) The conditioning σ-field $\mathcal{A}$.
   (ii) The probability measure $q_{\mathcal{A}}$ defined on $\mathcal{A}$.
   (iii) The conditional expectation $\mathcal{E}(\cdot \mid \mathcal{A})$.
3. If the σ-field $\mathcal{A}$ is generated by a countably infinite number of mutually orthogonal elements each having non-zero p-probability, then the corresponding $\mathcal{A}$-conditional expectation is of the form given by Eq. (2). In this case the $(\mathcal{A}, \psi_{\mathcal{A}})$-conditional probability measure q specified by Definition 5 is identical to the one obtained by using the method of "Jeffrey conditionalization" (Jeffrey 1965). Thus Jeffrey conditionalization is a special case of conditionalization via conditional expectation—although this connection does not seem to be well known [Gyenis and Rédei (2016) makes this connection more explicit].
4. Putting $Z = X$ in the defining property (ii) of the conditional expectation (Eq. (22)) and remembering that $p_{\mathcal{A}}$ is the restriction of p to $\mathcal{A}$, we obtain:

$$\int_X \mathscr{E}(\chi_B \mid \mathcal{A}) dp = \int_X \chi_B dp = p(B) \tag{27}$$

This requirement should be familiar: Eq. (27) is the "theorem of total probability". This becomes more transparent if one sees how it holds when $\mathcal{A}$ is a σ-field generated by a countable partition A_i $(i = 1, 2, \ldots)$ such that $p(A_i) \neq 0$ for every i. In this case we have (cf. Remark 2)

$$\mathscr{E}(\chi_B \mid \mathcal{A}) = \sum_i \frac{p(B \cap A_i)}{p(A_i)} \chi_{A_i} \tag{28}$$

So we can calculate

$$\int_X \mathscr{E}(\chi_B \mid \mathcal{A}) dp = \int_X \sum_i \frac{p(B \cap A_i)}{p(A_i)} \chi_{A_i} dp \tag{29}$$

$$= \sum_i \frac{p(B \cap A_i)}{p(A_i)} \int_X \chi_{A_i} dp \tag{30}$$

$$= \sum_i \frac{p(B \cap A_i)}{p(A_i)} p(A_i) \tag{31}$$

$$= \sum_i p(B \cap A_i) \tag{32}$$

$$= p(B) \tag{33}$$

Remark 4 The assumption of continuity of the linear functional $\psi_{\mathcal{A}}$ in the definition of Bayesan statistical inference and in the related definition of conditional probability (Definitions 4 and 5) entails that $q_{\mathcal{A}}$ is absolutely continuous with respect to the background measure p. Without the absolute continuity of $q_{\mathcal{A}}$ the linear functional $\psi_{\mathcal{A}} \circ \mathscr{E}(\cdot \mid \mathcal{A})$ on $\mathcal{L}^1(X, \mathcal{S}, p)$ is *not* an extension of $\psi_{\mathcal{A}}$ in general: If A is such that $p(A) = 0$ and $0 < q(A) < 1$ then $\mathscr{E}(\cdot \mid \mathcal{A})$ can happen to be a version of the conditional expectation such that $\mathscr{E}(\chi_A \mid \mathcal{A}) = \frac{1}{q(A)}\chi_{\mathcal{A}}$, which entails

$$q(A) = \psi_{\mathcal{A}}(\mathscr{E}(\chi_A \mid \mathcal{A})) = \frac{1}{q(A)} q(A) = 1 \neq q(A) \tag{34}$$

But even if q is not absolutely continuous with respect to p, the composition $\psi_{\mathcal{A}} \circ \mathscr{E}(\cdot \mid \mathcal{A})$ can still be an extension of $\psi_{\mathcal{A}}$ under some special circumstances. For instance if $p(A) = 0$ and $q_{\mathcal{A}}(A) = 1$, and $\mathscr{E}(\cdot \mid \mathcal{A})$ is a version such that $\mathscr{E}(\chi_A \mid \mathcal{A}) = \chi_{\mathcal{A}}$, then $\psi_{\mathcal{A}} \circ \mathscr{E}(\cdot \mid \mathcal{A})$ is an extension $\psi_{\mathcal{A}}$. In this case the conditional probability q depends also on the particular version of the conditional expectation used to extend $q_{\mathcal{A}}$. This situation occurs in the Borel–Kolmogorov Paradox situations as we will see in Sects. 3 and 4.

3 Conditional probabilities on probability zero events on the unit square calculated using conditional expectations

In this section we illustrate the notion of conditional expectation and conditional probabilities with respect to probability zero events defined in terms of conditional expectation by describing a simple example that is regarded in probability theory as paradigmatic.

Let (X, S, p) be the probability space with $X = [0, 1] \times [0, 1]$ the unit square in two dimension, S the Lebesgue measurable sets of $[0, 1] \times [0, 1]$ and p the Lebesgue measure on S. Let $C \doteq [0, 1] \times \{z\}$ be any horizontal slice of the square at number $z \in [0, 1]$ and $B \doteq b \times \{z\}$ be a Lebesgue measurable set of the square with b a measurable set in the slice (see the Fig. 1). What is the conditional probability of B on condition C? This question is the perfect analogue of the question asked in the Borel–Kolmogorov Paradox: the square replaces the sphere, C corresponds to a great circle and B to the arc on the circle. Furthermore, one may have the intuition that the answer to the question is determined: the conditional probability of B on condition C should be equal to the length $l(b)$ (one-dimensional Lebesgue measure) of b. But the "ratio analysis" (Bayes' rule) does not provide this answer because C has measure zero in the Lebesgue measure on the square. We have the square version of the Borel–Kolmogorov situation if we assume that the probability space on the square represents choosing a point randomly on the square.

Application of conditionalization via conditional expectation to this situation is the following. Consider the σ-field $\mathcal{A} \subset \mathcal{S}$ generated by the sets of form $[0, 1] \times A$ with A a Lebesgue measurable subset of $[0, 1]$. Note that $\mathcal{A}$ contains the slices $[0, 1] \times \{z\}$ where z is a number in $[0, 1]$; these sets have measure zero in the Lebesgue measure on the square. Then the $\mathcal{A}$-conditional expectation

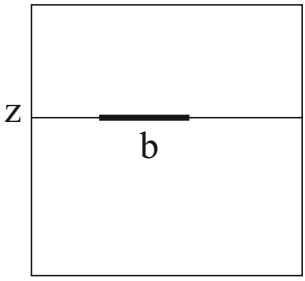


Fig. 1 The Borel–Kolmogorov Paradox situation on the unit square

$$\mathscr{E}(\cdot \mid \mathcal{A}) \colon \mathcal{L}^1([0,1] \times [0,1], \mathcal{S}, p) \to \mathcal{L}^1([0,1] \times [0,1], \mathcal{A}, p_{\mathcal{A}}) \tag{35}$$

exists, and an elementary calculation shows that the defining conditions (i) and (ii) in Definition 3 hold for the $\mathscr{E}(\cdot \mid \mathcal{A})$ given explicitly by:

$$\mathscr{E}(f \mid \mathcal{A})(x, y) = \int_0^1 f(x, y)dx \qquad \forall (x, y) \in [0,1] \times [0,1] \tag{36}$$

Inserting the characteristic function χ_B of $B = b \times \{z\}$ in the place of f in Eq. (36) one obtains for all $(x, y) \in [0,1] \times [0,1]$:

$$\mathscr{E}(\chi_B \mid \mathcal{A})(x, y) = \int_0^1 \chi_{b \times \{z\}}(x, y)dx \tag{37}$$

$$= \begin{cases} l(b), & \text{if } y = z \\ 0, & \text{if } y \neq z \end{cases} \tag{38}$$

If $q_{\mathcal{A}}$ is the probability measure on the σ-field $\mathcal{A}$ such that

$$q_{\mathcal{A}}(C) = q_{\mathcal{A}}([0,1] \times \{z\}) = 1 \tag{39}$$

$$q_{\mathcal{A}}(C^{\perp}) = q_{\mathcal{A}}(([0,1] \times \{z\})^{\perp}) = 0 \tag{40}$$

then, by the definition of Bayesian statistical inference (see also Remark 4), the $(\mathcal{A}, q_{\mathcal{A}})$-conditional probability $q(b \times \{z\}))$ of B on condition $C = [0,1] \times \{z\}$, i.e. on condition that $q_{\mathcal{A}}([0,1] \times \{z\}) = 1$, can be calculated using (37):

$$q(b \times \{z\}) = q_{\mathcal{A}}(\mathscr{E}(\chi_{b \times \{z\}} \mid \mathcal{A})) \tag{41}$$

$$= \int_{[0,1] \times [0,1]} \mathscr{E}(\chi_{b \times \{z\}} \mid \mathcal{A}) dq_{\mathcal{A}} \tag{42}$$

$$= l(b) \tag{43}$$

This is in complete agreement with intuition: Given any one dimensional slice $C = [0,1] \times \{z\}$ at point z across the square, the $(\mathcal{A}, q_{\mathcal{A}})$-conditional probability of the subset b of that slice on condition that we are on that slice ($q_{\mathcal{A}}(C) = 1$) is proportional to the length of the subset b. This result is obtained using the technique of conditional

expectation with respect to a sub-σ-field $\mathcal{A}$ some elements of which have probability zero. This is regarded as a classic example of conditioning with respect to probability zero events (Billingsley 1995, p. 432).

The phenomenon of the conditional expectation being determined only up to a probability zero set also can be illustrated on this example. We know that conditional expectations are defined up to measure zero only (Proposition 2). Thus, the conditional expectation $\mathcal{E}(\cdot \mid \mathcal{A})$ defined by (36) is just one *version* of the conditional expectation determined by the σ-field $\mathcal{A}$. Another version $\mathscr{E}_m(\cdot \mid \mathcal{A})$ of the $\mathcal{A}$-conditional expectation can be obtained by choosing a particular $z_0 \in [0, 1]$ and defining $\mathscr{E}_m(\cdot \mid \mathcal{A})$ by

$$\mathscr{E}_m(f \mid \mathcal{A})(x, y) \doteq \begin{cases} \mathscr{E}(f \mid \mathcal{A})(x, y), & \text{if } y \neq z_0 \\ \int_0^1 \rho(x) f(x, y) dx, & \text{if } y = z_0 \end{cases} \tag{44}$$

where ρ is a probability density function for a probability measure m on $[0, 1]$ (with respect to the Lebesgue measure on $[0, 1]$). Computing the conditional probability $q(b \times \{z_0\})$ along the lines of (41)–(43) using the $\mathscr{E}_m(\cdot \mid \mathcal{A})$ version of the $\mathcal{A}$-conditional expectation one obtains

$$\begin{aligned} q(b \times \{z_0\}) &= \psi_{\mathcal{A}}(\mathscr{E}_m(\chi_{b \times \{z_0\}} \mid \mathcal{A})) && (45) \\ &= \int_{[0,1] \times [0,1]} \mathscr{E}_m(\chi_{b \times \{z_0\}} \mid \mathcal{A}) dq_{\mathcal{A}} && (46) \\ &= \int_b \rho(x) dx && (47) \\ &= m(b) && (48) \end{aligned}$$

Thus, given $\mathcal{A}$ and *any, fixed*, one dimensional slice of the square, one obtains different values for the conditional probability of Lebesgue measurable subsets of that slice depending on which version of the $\mathcal{A}$-conditional expectation one uses to calculate it. Using the "canonical" version given by (36) one obtains the value proportional to the length, using the m-version $\mathscr{E}_q(\cdot \mid \mathcal{A})$ given by (44) one obtains the value $m(b)$. Fixing the σ-field alone does not determine any of these two versions, or indeed any of an uncountable number of other versions, in harmony with the conditional expectation being undetermined up to a measure zero set. But then what singles out the canonical version?

Having a look at the definition of $\mathscr{E}(\cdot \mid \mathcal{A})$ [Eq. (44)], one realizes that it is the particular mathematical structure of the situation that makes that definition possible and thus singles out the canonical version: the set of elementary events of the probability space on the unit square has the form of product $[0, 1] \times [0, 1]$, and one can perform a partial integral with respect to one variable in the product probability space. These two conditions together with the specific form and location of the conditioning σ-field in the product structure determine not only a conditional expectation that yields the "proportional-to-the-length" value $l(b)$ on all slices except for sets of slices that have measure zero in the two dimensional Lebesgue measure but a version that yields the "intuitively right" conditional probabilities on *every* slice.

The crucial role of the product structure in the existence of the canonical version of the conditional expectation can also be seen if one realizes that the reasoning involving Eqs. (35)–(43) remain valid without any change if one replaces (i) the unit square with the Lebesgue measure on it by any product space $(X_1 \times X_2, \mathcal{S}_1 \otimes \mathcal{S}_2, p_1 \times p_2)$, and (ii) the σ-field $\mathcal{A}$ by a σ-field generated by elements of the form $X_1 \times B$ ($B \in \mathcal{S}_2$). Hence, even if the component probability spaces $(X_1, \mathcal{S}_1, p_1)$ and $(X_2, \mathcal{S}_2, p_2)$ in the product have finite Boolean algebras (and consequently so does the product space), and even if some events in the component algebras have probability zero, the analogue of the canonical conditional expectation (35) exists and yields probabilities conditional on probability zero events via the analogue of Eq. (43), although it is very clear that conditional expectations are genuinely undetermined on probability zero events in finite probability spaces in general. To see this finite situation explicitly, consider the following simple example:

Let $(X_1, \mathcal{S}_1, p_1)$ be generated by two atomic events B_1 and B_2 and $(X_2, \mathcal{S}_2, p_2)$ be generated by two atomic events C_1 and C_2 with probabilities

$$p_1(B_1) = \frac{1}{2} \qquad p_1(B_2) = \frac{1}{2} \tag{49}$$

$$p_2(C_1) = 0 \qquad p_2(C_2) = 1 \tag{50}$$

Then the product space

$$(X_1 \times X_2, \mathcal{S}_1 \otimes \mathcal{S}_2, p_1 \times p_2) \tag{51}$$

is generated by the four atomic events

$$A_1 = B_1 \times C_1;\ A_2 = B_2 \times C_1;\ A_3 = B_1 \times C_2;\ A_4 = B_2 \times C_2 \tag{52}$$

($p_1 \times p_2$ is the product measure). Let $\mathcal{C}$ be the Boolean algebra generated by elements of the form $X_1 \times C$ ($C \in \mathcal{S}_2$) and $p_{\mathcal{C}}$ be the restriction of $p_1 \times p_2$ to $\mathcal{C}$. Then there exists a conditional expectation

$$\mathscr{E}(\cdot \mid \mathcal{C})\colon \mathcal{L}^1(X_1 \times X_2, \mathcal{S}_1 \otimes \mathcal{S}_2, p_1 \times p_2) \to \mathcal{L}^1(X_1 \times X_2, \mathcal{C}, p_{\mathcal{C}}) \tag{53}$$

given by

$$\mathscr{E}(f \mid \mathcal{C})(x_i, y_j) = \frac{1}{2} \sum_i f(x_i, y_j) \qquad x_i = B_i,\ y_j = C_j \quad (i, j = 1, 2) \tag{54}$$

which we call the *canonical* conditional expectation. Using the delta function notation

$$\delta_{i,j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \tag{55}$$

the values of the characteristic function $\chi_{B_1 \times C_1}$ can be written as

$$\chi_{B_1 \times C_1}(x_i, y_j) = \delta_{1i}\delta_{1j} \qquad x_i = B_i,\ y_j = C_j \quad (i, j = 1, 2) \tag{56}$$

and so the value of the canonical conditional expectation on the characteristic function $\chi_{B_1\times C_1}$ can be computed explicitly:

$$\mathscr{E}(\chi_{B_1\times C_1} \mid \mathcal{C})(x_i, y_j) = \frac{1}{2}\delta_{1j} \qquad x_i = B_i,\ y_j = C_j \quad (i, j = 1, 2) \tag{57}$$

If $q_{\mathcal{C}}$ is the probability measure on $\mathcal{C}$ such that

$$q_{\mathcal{C}}(X_1 \times C_1) = 1 \tag{58}$$
$$q_{\mathcal{C}}(X_1 \times C_2) = 0 \tag{59}$$

then the canonical conditional expectation (54) yields a definite conditional probability $q(B_1 \times C_1)$ on condition $X_1 \times C_1$, which is a measure zero event in the probability space $(X_1 \times X_2, \mathcal{S}_1 \otimes \mathcal{S}_2, p_1 \times p_2)$:

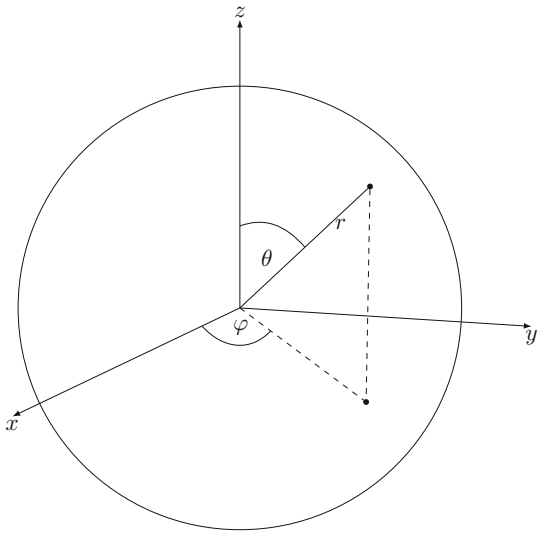
$$q(B_1 \times C_1) = q_{\mathcal{C}}(\mathscr{E}(\chi_{B_1\times C_1} \mid \mathcal{C})) = \tag{60}$$
$$= \int_{X_1\times X_2} \mathscr{E}(\chi_{B_1\times C_1} \mid \mathcal{C}) dq_{\mathcal{C}} \tag{61}$$
$$= \sum_j \frac{1}{2}\delta_{1j}\, q_{\mathcal{C}}(X_1 \times C_j) = \frac{1}{2} q_{\mathcal{C}}(X_1 \times C_1) = \frac{1}{2} \tag{62}$$

Thus, using the technique of conditional expectation in a probability space with a *finite* Boolean algebra, we have obtained a definite value for the conditional probability of an event with respect to a conditioning event that has probability zero. On the other hand, we know that in finite Boolean algebras conditional expectations are always of the form (24) (cf. Remark 2), and that form clearly shows that the value of the conditional expectation is *not determined* on probability zero events in general; hence conditional probabilities with respect to probability zero events are also left undetermined. There is of course no contradiction here. The point is that the product structure singles out a *particular version* of the conditional expectation with respect to a particular Boolean subalgebra that is located in a specific position with respect to the product so that a version of the conditional expectation can be obtained by taking a partial integral. This version in turn yields a specific value for conditional probabilities with respect to probability zero events. It is important to emphasize that the product structure just *singles out* the canonical version but does not entail it logically because any version is compatible with the product structure.

4 "Intuitively correct" conditional probabilities with respect to probability zero events in the Borel–Kolmogorov Paradox obtained using conditional expectations

Consider now the probability space $(S, \mathcal{B}(S), p)$ on the unit sphere S in three dimension with the surface measure p on the Lebesgue sets $\mathcal{B}(S)$ on S. Choose a great circle C on S. We wish to calculate the conditional probability of an arc B on condition that

Fig. 2 Polar coordinates



the arc is on the great circle C. One can calculate this conditional probability following exactly the steps used to calculate the conditional probability of the subset b of a slice of the square on condition that b is on that slice. The only difference is in the slight complication due to the non-trivial geometry of the sphere.

Points of the unit sphere S can be given by polar coordinates:

$$S = \left\{(x(\phi, \theta, r), y(\phi, \theta, r), z(\phi, \theta, r)) : \ 0 \leq \phi \leq 2\pi, \ 0 \leq \theta \leq \pi, \ r = 1\right\} \quad (63)$$

where

$$x(\phi, \theta, r) = r \cos\phi \sin\theta \quad (64)$$

$$y(\phi, \theta, r) = r \sin\phi \sin\theta \quad (65)$$

$$z(\phi, \theta, r) = r \cos\theta \quad (66)$$

(See Fig. 2)

In our case $r = 1$ is fixed whence each mapping $f(x, y, z) : S \to \mathbb{R}$ can be identified with a two-variable function

$$f(\phi, \theta) = f\big(x(\phi, \theta, 1), y(\phi, \theta, 1), z(\phi, \theta, 1)\big) : [0, 2\pi] \times [0, \pi] \to \mathbb{R} \quad (67)$$

and the sphere S and the chosen great circle C can be identified with the sets

$$S = \{(\phi, \theta) \ : \ \phi \in [0, 2\pi], \theta \in [0, \pi]\} = [0, 2\pi] \times [0, \pi] \quad (68)$$

$$C = \{(\phi, \pi/2) \ : \ \phi \in [0, 2\pi]\} = [0, 2\pi] \times \{\pi/2\} \quad (69)$$

Let $\mathcal{O}$ be the σ-field generated by measurable sets of circles on the sphere plane of which is parallel to that of the chosen great circle C (see Fig. 3):

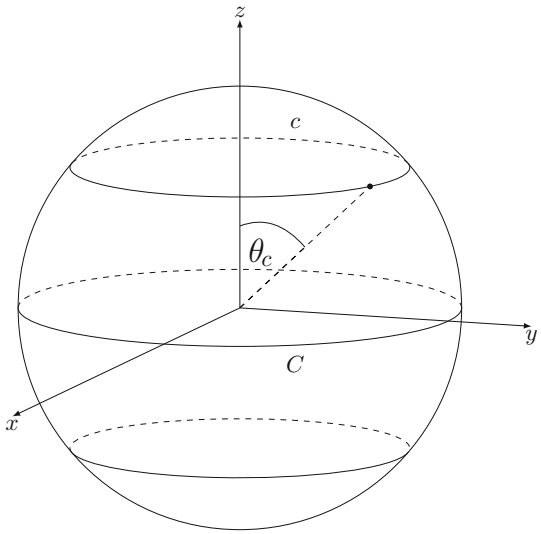


Fig. 3 Parallel circles generating σ-algebra $\mathcal{O}$

$$\mathcal{O} = \big\{[0, 2\pi] \times A : \ A \subseteq [0, \pi] \text{ is measurable}\big\} \tag{70}$$

There exist then the $\mathcal{O}$-conditional expectation

$$\mathcal{E}(\cdot \mid \mathcal{O}) \colon \mathcal{L}^1(S, \mathcal{B}(S), p) \to \mathcal{L}^1(S, \mathcal{O}, p_{\mathcal{O}}) \tag{71}$$

and one can verify (see Appendix 1 in the Appendix section for details) that $\mathcal{E}(\cdot \mid \mathcal{O})$ given by

$$\mathcal{E}(f \mid \mathcal{O})(\phi, \theta) = \frac{1}{2\pi} \int_0^{2\pi} f(\phi, \theta)\, d\phi \tag{72}$$

is a version of the $\mathcal{O}$-conditional expectation.

Let χ_B be the characteristic function of an arc B on the great circle C specified by the angles ϕ_1 and ϕ_2:

$$B = [\phi_1, \phi_2] \times \{\pi/2\} \tag{73}$$

Then we have

$$\mathcal{E}(\chi_B \mid \mathcal{O})(\phi, \theta) = \frac{1}{2\pi} \int_0^{2\pi} \chi_B(\phi, \theta)\, d\phi \ = \ \frac{1}{2\pi} \int_{\phi_1}^{\phi_2} 1\, d\phi \tag{74}$$

$$= \frac{1}{2\pi} \begin{cases} \phi_2 - \phi_1 & \text{if } \theta = \frac{\pi}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{75}$$

If $q_{\mathcal{O}}$ is the probability measure on the sphere taking value 1 on the great circle C and value 0 on its complement $C^{\perp}$ (see also Remark 4), then the $(\mathcal{O}, q_{\mathcal{O}})$-conditional probability $q(B)$ of the arc B can be computed easily using (75)

$$q(B) = \psi_{\mathcal{O}}(\chi_B) = \int_S \mathcal{E}(\chi_B|\mathcal{O})dq_{\mathcal{O}} \tag{76}$$

$$= \frac{\phi_2 - \phi_1}{2\pi} \tag{77}$$

That is to say, the $(\mathcal{O}, q_{\mathcal{O}})$-conditional probability $q(B)$ of the arc B is proportional to the length of the arc. Thus, just like in case of the square, choosing a suitable sub-σ-field of the Lebesgue sets of the sphere, and using the device of conditional expectations defined by the chosen sub-σ-field, one can obtain the sought after conditional probabilities with respect to probability zero events in the Borel–Kolmogorov situation, and the calculated conditional probabilities are the "intuitively correct" ones. What is the problem then?

5 Conditional probability with respect to probability zero events in the Borel–Kolmogorov situation depends on the conditioning algebra

The alleged problem is that the conditional probabilities so obtained depend on the σ-field $\mathcal{O}$: if, instead of $\mathcal{O}$, one takes the σ-field $\mathcal{M}$ generated by (measurable sets of) great circles that intersect C at the same two points ("meridian circles" with respect to *North* and *South Poles*), then the $(\mathcal{M}, q_{\mathcal{M}})$-conditional probability of the arc B will be different from the $(\mathcal{O}, q_{\mathcal{O}})$-conditional probability of the arc B: One can calculate these $(\mathcal{M}, q_{\mathcal{M}})$-conditional probabilities of B following exactly the steps in the preceding Sect. 4 which led to the $(\mathcal{O}, q_{\mathcal{O}})$-conditional probabilities: Choose a meridian circle C that in the introduced polar coordinates is given by

$$C = \{(0, \theta), (\pi, \theta) \ : \ \theta \in [0, \pi]\} = \{0, \pi\} \times [0, \pi] \tag{78}$$

That is to say: C is the meridian circle at longitude 0 (see Fig. 4). Call a set $A \subseteq [0, 2\pi]$ symmetric if $x \in A$ implies that $(x + \pi)$ modulo 2π also belongs to A. Note that $\{0, \pi\}$ in (78) is symmetric, and each collection of meridian circles correspond to a set of form $A \times [0, \pi]$, where A is symmetric. (Remark: one could get rid of requiring symmetry by letting the parameter θ run from 0 to 2π). Let $\mathcal{M}$ be the σ-field generated by all measurable sets of meridian circles:

$$\mathcal{M} = \left\{A \times [0, \pi] : \ A \subseteq [0, 2\pi] \text{ is measurable and symmetric}\right\} \tag{79}$$

There exist then the $\mathcal{M}$-conditional expectation

$$\mathcal{E}(\cdot \mid \mathcal{M}) \colon \mathcal{L}^1(S, \mathcal{B}(S), p) \to \mathcal{L}^1(S, \mathcal{M}, p_{\mathcal{M}}) \tag{80}$$

One can verify (for details see Appendix 2 in the Appendix section) that a version of $\mathcal{E}(\cdot \mid \mathcal{M})$ is given by

$$\mathcal{E}(f \mid \mathcal{M})(\phi, \theta) = \frac{1}{2} \int_0^{2\pi} f(\phi, \theta)|\sin\theta| \, d\theta \tag{81}$$

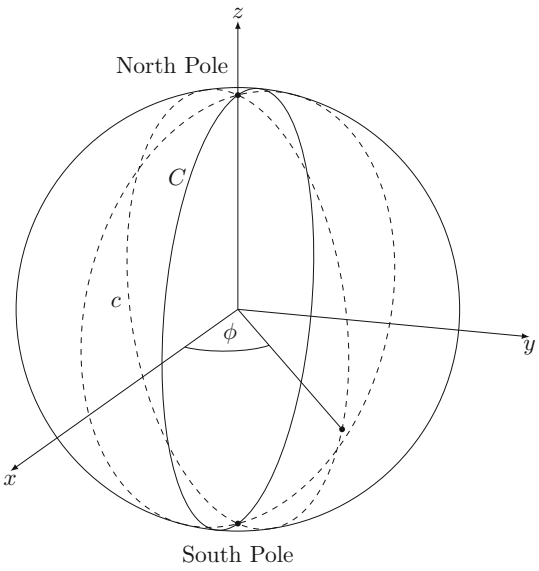


Fig. 4 Meridian circles generating σ-field $\mathcal{M}$

Let χ_B be the characteristic function of an arc on the meridian circle C specified by the angles $0 \leq \theta_1 < \theta_2 \leq \pi$:

$$B = \{0\} \times [\theta_1, \theta_2] \tag{82}$$

Then we have

$$\mathcal{E}(\chi_B \mid \mathcal{M})(\phi, \theta) = \frac{1}{2} \int_0^{2\pi} \chi_B(\phi, \theta) |\sin\theta| \, d\theta \tag{83}$$

$$= \frac{1}{2} \begin{cases} \int_{\theta_1}^{\theta_2} \sin\theta \, d\theta = \cos\theta_1 - \cos\theta_2 & \text{if } \phi = 0 \\ 0 & \text{otherwise} \end{cases} \tag{84}$$

If $q_{\mathcal{M}}$ is the probability measure on the σ-field $\mathcal{M}$ taking value 1 on the meridian circle C and value 0 on its complement $C^{\perp}$ (see also Remark 4), then the $(\mathcal{M}, q_{\mathcal{M}})$-conditional probability $q(B)$ of the arc B can be calculated easily by using (83):

$$q(B) = \psi_{\mathcal{M}}(\chi_B) = \int_S \mathcal{E}(\chi_B \mid \mathcal{M}) \, dq_{\mathcal{M}} \tag{85}$$

$$= \frac{\cos\theta_1 - \cos\theta_2}{2} \tag{86}$$

Clearly, the $(\mathcal{M}, q_{\mathcal{M}})$-conditional (86) and $(\mathcal{O}, q_{\mathcal{O}})$-conditional (77) probabilities are different.

Note that, just like in case of the square, both the $\mathcal{O}$-conditional and the $\mathcal{M}$-conditional expectation are only determined only up to probability zero events, and the

definitions (72) and (81) yield specific versions of the respective conditional expectations. These versions are singled out, again, by the fact that the sphere and the circles on it have a Cartesian product structure and the conditioning σ-fields are located in a specific position with respect to the product so that a version of the conditional expectation can be obtained by taking a partial integral.

Remark 5 The $\mathcal{O}$-conditional expectation (72) can be used to calculate the $\mathcal{O}$-conditional probability on any circle c parallel to the great circle C specified by (69): the calculation following the steps (73)–(77) results in a uniform distribution on any such circle c. Similarly: the $\mathcal{M}$-conditional expectation (81) can be used to calculate the $\mathcal{M}$-conditional probability on any meridian circle C_M replacing the great circle C specified by (78). Repeating the steps (82)–(86) one obtains the $\mathcal{M}$-conditional distribution (86) on C_M.

Given a great circle C one could of course consider the four element σ-field $\mathcal{A}$ containing C, its complement, the empty set and the whole sphere, and compute the $(\mathcal{A}, q_{\mathcal{A}})$-conditional probability of an arc on C, using the $\mathcal{A}$-conditional expectation. Since this $\mathcal{A}$ is generated by a countable set of disjoint elements, we know (Remark 2) what form the $\mathcal{A}$-conditional expectation has in this case, and we also know that since C has measure zero in the surface measure of the sphere, the value of the $\mathcal{A}$-conditional expectation on C is left undetermined. Thus we can take any value we regard as "intuitively correct", and choose the corresponding version of the $\mathcal{A}$-conditional expectation. Thus conditionalizing using the theory of conditional expectations can accommodate any value of conditional probability on a probability zero event, including the "intuitively correct" uniform conditional probability. But this conditional probability is not determined in the theory of conditionalization by choosing the conditioning algebra to be $\mathcal{A}$ and by the stipulation that the probability on the sphere is given by the uniform measure. We will explain in Sect. 7 why one may have the *wrong* intuition that it is.

6 Is dependence of conditional probability on the conditioning algebra paradoxical?

One finds in the literature two types of worries concerning the $(\mathcal{M}, q_{\mathcal{M}})$-conditional and $(\mathcal{O}, q_{\mathcal{O}})$-conditional probabilities. One is what we call, using Howson's terminology (Howson 2014, p. 8), the *Description-Relativity Worry*, the other is that the $(\mathcal{M}, q_{\mathcal{M}})$-conditional probability is counterintuitive. These two worries form the heart of the Borel–Kolmogorov Paradox. In this section we will show that the Description-Relativity Worry rests on a misinterpretation of what the difference between the $(\mathcal{M}, q_{\mathcal{M}})$-conditional and $(\mathcal{O}, q_{\mathcal{O}})$-conditional probabilities signify, and in Sect. 7 we will argue that the $(\mathcal{M}, q_{\mathcal{M}})$-conditional probabilities are not counterintuitive.

The Description-Relativity Worry is the concern that when it comes to calculate any probability, conditional probabilities included, it should not matter how the random events involved are described: what specific parameters are used to refer to random events and what coordinate system is used to fix a particular notation in which probabilistic calculations are carried out should be a matter of convention, not

affecting the values of probabilities. In what follows we use the general term "labeling" to refer to any description, parametrization, coordinatization etc. of random events. The Description-Relativity Worry is then that the $(\mathcal{M}, q_{\mathcal{M}})$-conditional and $(\mathcal{O}, q_{\mathcal{O}})$-conditional probabilities violate what one can call "Labeling Irrelevance": the norm that values of probabilities should not depend on labeling. This is a very important principle, which is crucial in probabilistic modeling: its violation is not compatible with an objective interpretation of probability (this is argued in detail in Gyenis and Rédei (2015a), where it is shown that Bertrand's Paradox does not entail violation of Labeling Irrelevance). Subjective interpretations of probability are a different matter: a subject's degrees of beliefs might depend on particular labeling of random events, as Rescorla argues Rescorla (2014) [see also Easwaran (2008)]. We do not wish to discuss this situation, see the end of Sect. 7 for some brief comments. In any case, it is obviously important to know whether the Description-Relativity Worry is indeed justified in connection with the difference of the $(\mathcal{M}, q_{\mathcal{M}})$-conditional and $(\mathcal{O}, q_{\mathcal{O}})$-conditional probabilities. We claim it is not.

Rescorla derives the conditional probabilities (77) and (86) using the technique of calculating conditional probability density functions (pdf's) rather than specifying the two σ-fields $\mathcal{O}$ and $\mathcal{M}$ explicitly and calculating the respective conditional expectations. Having done this, Rescorla expresses the Description-Relativity Worry thus:

> ... conditional probability density is not invariant under coordinate transformation. Standard techniques for computing conditional pdfs yield different solutions, depending upon our choice of coordinates. Apparently, then, the coordinate system through which I describe a null event impacts how I should condition on the null event. This dependence upon a coordinate system looks paradoxical. Since the null event remains the same, shouldn't I obtain the same answer either way? (Rescorla 2014, p. 10)

Viewing the difference between the $(\mathcal{M}, q_{\mathcal{M}})$-conditional (86) and $(\mathcal{O}, q_{\mathcal{O}})$-conditional (77) probabilities as violation of Labeling Irrelevance is however a misinterpretation of the phenomenon. This becomes transparent when one specifies more carefully what "coordinate transformation", "different descriptions" or "re-labeling" of random events are.

Assume that $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ are probability spaces. Then $(X', \mathcal{S}')$ can be viewed as a re-labeled copy of $(X, \mathcal{S})$ if there exists a bijection f between X and X' such that both f and its inverse f^{-1} are measurable: the inverse image under f of every $A' \in \mathcal{S}'$ is in $\mathcal{S}$, and the inverse image under the inverse function f^{-1} of every $A \in \mathcal{S}$ is in $\mathcal{S}'$. The function f is then called a re-labeling. Note that without the double-measurability condition the function f cannot be considered a re-labeling because if the inverse function f^{-1} were not measurable, then some elements in $\mathcal{S}$ would be "lost" when passing via f from $(X, \mathcal{S})$ to $(X', \mathcal{S}')$: there would then exist an $A \in \mathcal{S}$ such that $f[A] = \{f(x) : x \in A\} \notin \mathcal{S}'$. Similarly: if f were not measurable, then there would be an element $A' \in \mathcal{S}'$ that refers to some general random event that is part of the phenomenon $(X', \mathcal{S}', p')$ is a model of, but $f^{-1}[A'] = \{f^{-1}(x') : x' \in A'\} \notin \mathcal{S}$, hence under the re-labeling f that random event would be lost in the model $(X, \mathcal{S}, p)$. In this case the two probability spaces $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ obviously could not

be regarded as models of the *same* random phenomenon with the only difference that random events are differently labeled in them. Because of the double measurability condition on re-labeling f, a re-labeling gives rise to an isomorphism h_f between the σ-fields $\mathcal{S}$ and $\mathcal{S}'$ (h_f is the inverse image function of the inverse function f^{-1} of f).

Recall that if f is a re-labeling between X and X', and f and f^{-1} also preserve p and p', respectively, in the sense that (87)–(88) below hold

$$p'(f[A]) = p(A) \quad \text{for all} \quad A \in \mathcal{S} \tag{87}$$

$$p(f^{-1}[A']) = p'(A') \quad \text{for all} \quad A' \in \mathcal{S}' \tag{88}$$

then the probability spaces $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ are called isomorphic as probability spaces and f is a (measure theoretic) isomorphism (Aaronson 1997, p. 3). It is obvious that a re-labeling need not be a measure theoretic isomorphism in general. Less obvious is that a re-labeling is not necessarily a measure theoretic isomorphism even if the probability measures are very special; possibly so special that one expects re-labelings to be isomorphisms: this happens when p and p' are both Haar measures. This lies at the heart of Bertrand's Paradox, see Gyenis and Rédei (2015a, b) for details.

Labeling Irrelevance can now be expressed by the claim that when describing a phenomenon probabilistically, we can choose either the $(X, \mathcal{S})$ or the $(X', \mathcal{S}')$ labeling of random events as long as there is a re-labeling f between X and X'. Indeed: nothing can prevent us choosing either from elements of $(X, \mathcal{S})$ or from elements of $(X', \mathcal{S}')$ when we wish to refer to random events, and if we choose $(X, \mathcal{S})$, then we can specify a probability measure p on $\mathcal{S}$ such that the probability space $(X, \mathcal{S}, p)$ is a good model of the phenomenon. Choosing the probability $p'[A'] \doteq p(f^{-1}[A'])$ on $(X', \mathcal{S}')$ makes $(X', \mathcal{S}', p')$ also a good model of the phenomenon and $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ will be isomorphic as probability spaces with respect to f. In short Labeling Irrelevance, the conventionality of labeling of random events in probabilistic modeling, is expressed by the claim that measure theoretically isomorphic probability spaces can be used to describe the same random phenomenon.

An example of re-labeling is passing from the Cartesian coordinates to the polar coordinates when describing the sphere and its measurable subsets: the transformation (64)–(66) is a double measurable bijection. Any point and any measurable subset on the sphere can be expressed either in the (x, y, z) coordinates or in the (ϕ, θ, r) coordinates.

It should now be clear that the difference between the $(\mathcal{M}, q_{\mathcal{M}})$-conditional (86) and $(\mathcal{O}, q_{\mathcal{O}})$-conditional (77) probabilities is not a case of violation of Labeling Irrelevance: the two conditional probabilities cannot be considered as conditional probabilities of the *same* event with respect to the *same* conditioning conditions in *different* "co-ordinatizations" (labelings): When one calculates the conditional probabilities of an event $A \in \mathcal{S}$ in a different, "primed" labeling (in $\mathcal{S}'$), then the conditioning conditions also have to be considered in the primed labeling, otherwise the conditioning is not with respect to the same conditions. Thus if $\mathcal{A}$ is a sub-σ-field of $\mathcal{S}$ and one computes the $\mathcal{A}$-conditional expectation in $(X, \mathcal{S}, p)$ and the corresponding $\mathcal{A}$-conditional probabilities of A, then to obtain the conditional probabilities of the *same* event in the primed labeling with respect to the *same* conditioning conditions, calculated in $(X', \mathcal{S}', p')$, one has to use the $h_f(\mathcal{A})$-conditional expectation in $(X', \mathcal{S}', p')$

to calculate the conditional probabilities of $h_f(A)$. Here h_f is the Boolean algebra isomorphism between $\mathcal{S}$ and $\mathcal{S}'$ determined by the re-labeling f. The restriction of h_f to $\mathcal{A}$ is a Boolean algebra isomorphism between $\mathcal{A}$ and $h_f(\mathcal{A})$ and so the $\mathcal{A}$-conditional probability of A and the $\mathcal{A}'$-conditional probability of A' can be regarded as the conditional probability of the *same* event in different labeling with respect to the *same* conditions in different labeling only if there exists a Boolean algebra isomorphism h between $\mathcal{A}$ and $\mathcal{A}'$ such that $h(A) = A'$.

There exists however no Boolean algebra isomorphism h between the σ-field $\mathcal{O}$ generated by the measurable sets of circles parallel to a great circle C and the σ-field $\mathcal{M}$ generated by the measurable sets of meridian circles such that $h(C)$ is a great (meridian) circle in $\mathcal{M}$. This can be seen by a simple indirect proof: Assume the contrary, i.e. that h is an $\mathcal{O} \to \mathcal{M}$ Boolean algebra isomorphism, C is the great circle in $\mathcal{O}$, and $h(C)$ is a great circle in $\mathcal{M}$. The circles c in $\mathcal{O}$ parallel to the great circle C are the atoms of $\mathcal{O}$ and these are the only atoms in $\mathcal{O}$. The atomic structure of a σ-field is preserved under isomorphism, so $h(c)$ are the (only) atoms in $\mathcal{M}$. Since the two element set {*North Pole*, *South Pole*} is an atom in $\mathcal{M}$, there is a $c_0 \in \mathcal{O}$ such that $h(c_0) = \{$*North Pole*, *South Pole*$\}$; furthermore $c_0 \neq C$ because $h(C)$ is assumed to be a great circle. We have $C \cap c = \emptyset$ for any circle $c \in \mathcal{O}$ parallel to C and different from C, in particular $C \cap c_0 = \emptyset$, which entails (h being an isomorphism)

$$\emptyset = h(C \cap c_0) \tag{89}$$
$$= h(C) \cap h(c_0) = h(C) \cap \{\textit{North Pole}, \textit{South Pole}\} \tag{90}$$
$$= \{\textit{North Pole}, \textit{South Pole}\} \tag{91}$$

(the last equation holding because $h(C)$ was assumed to be a great circle in $\mathcal{M}$, and all meridian circles contain both the *South* and the *North Poles*). Since (89)–(91) is a contradiction, no such isomorphism exists.

In fact, more is true: there exists no isomorphism between the subalgebras $\mathcal{O}$ and $\mathcal{M}$ *at all*. To see this, let $s = (\phi_0, \theta_0)$ be a point on the sphere such that $(\theta_0 \neq 0, \pi)$. We claim that the following are true:

(i) If $s \in A \in \mathcal{O}$, then the whole circle $\{(\phi, \theta_0) : \phi \in [0, 2\pi]\}$ parallel to the equator must be a subset of A.
(ii) If $s \in B \in \mathcal{M}$, then the meridian circle $\{(\phi_0, \theta) : \theta \in [0, \pi]\}$ has to be a subset of B.

(i) and (ii) can be proved by induction: the statements obviously hold for the generator elements of the algebras $\mathcal{O}$ and $\mathcal{M}$, and it is not hard to see that (i) and (ii) remain true under taking arbitrary unions, meets and complement. (ii) entails that the intersection of two non-disjoint elements $A, B \in \mathcal{M}$ must contain the set {*North Pole*, *South Pole*} (which belongs to $\mathcal{M}$). In other words, there is an element $C \neq \emptyset$ in $\mathcal{M}$ (namely $C = \{$*North Pole*, *South Pole*$\}$) such that for any two sets $A, B \in \mathcal{M}$, if $A \cap B \neq \emptyset$, then $C \subseteq A \cap B$. The same does not hold in $\mathcal{O}$: Let $A = \{c_{\theta_1}, c_{\theta_2}\}$ and $B = \{c_{\theta_1}, c_{\theta_3}\}$ be two sets of parallel circles with latitudes θ_1, θ_2 and θ_3. Then $A \cap B = \{c_{\theta_1}\}$. Taking two similar sets A' and B' of parallel circles one has $A' \cap B' = \{c_{\theta'_1}\}$ and clearly $c_{\theta_1} = c_{\theta'_1}$ need not hold, and this prevents the existence of an Boolean algebra isomorphism between $\mathcal{O}$ and $\mathcal{M}$.

Thus indeed "It can't be the case that, conditional on the chosen point lying on the circle that is the great circle containing the Greenwich meridian of our first coordinatization and is the equator of our second, we have different conditional distributions depending on how we describe the circle." (Myrvold 2014, p. 14) But there is no danger of such a counterintuitive dependence to occur in the Borel–Kolmogorov Paradox situation. This is because considering the great circle first as the "Greenwich meridian" in the σ-field $\mathcal{M}$ generated by meridian circles, and, second, as the "equator" element in the σ-field $\mathcal{O}$ are not "different descriptions" of the same great circle in two coordinatizations: Given a coordinatization (e.g. in terms of the polar coordinates), one can describe the circle uniquely as a particular set of ordered pairs of real numbers [see (69)]. Given Cartesian coordinates, one also can describe the same great circle as ordered pairs of different real numbers. *These* sets of pairs of numbers are different descriptions of the same great circle. When one considers the same great circle as an element in the σ-fields $\mathcal{O}$ and $\mathcal{M}$, respectively, and calculates the conditional probability distribution on the great circle using (particular versions of the) $\mathcal{O}$- and $\mathcal{M}$-conditional expectations, then one does not "re-coordinatize" or "re-describe" the great circle but calculates conditional probabilities with respect to different conditioning σ-fields. Each of these two conditional probabilities are invariant with respect to coordinatization (description) when re-coordinatization and re-description are properly understood as re-labelings that are measure theoretic isomorphisms. That these conditional probabilities are different is perfectly understandable and acceptable because they do not indicate a paradoxical dependence of conditional probabilities of the *same event* with respect to the *same conditioning* conditions in *different co-ordinatization* but a sensitive dependence of conditional probabilities of the *same event* on *different conditioning* σ-fields with respect to which conditional probabilities are defined in terms of conditional expectations. This latter dependence is however not only not paradoxical but entirely natural and expected once the concept of conditional probability is defined properly in terms of conditional expectations.

7 Why one may think that only the uniform conditional probability on a great circle in the Borel–Kolmogorov Paradox is correct

The conclusion of the previous section already indicates what we would like to formulate here explicitly: Both the $(\mathcal{O}, q_{\mathcal{O}})$-conditional and the $(\mathcal{M}, q_{\mathcal{M}})$-conditional distributions on the great circle are intuitively correct—when one has the correct concept of conditionalization in mind. To see this and to understand why one might have the intuition that only the uniform probability on any great circle is the intuitively correct conditional probability, one has to draw a clear distinction between (i) probability theory taken in itself as part of pure mathematics and (ii) probability theory as mathematical model of some phenomena (application of probability theory). The importance of these distinctions were emphasized in Gyenis and Rédei (2015a), where conceptual confusions resulting from disregarding them is analyzed from the perspective of another alleged paradox involving probability theory (Bertrand's Paradox).

7.1 Probability theory as pure mathematics

Probability theory taken in itself and defined by the Kolmogorovian axioms is part of pure mathematics, a branch of measure theory. A mathematical statement, claim, inference, is therefore probabilistic only if it can be stated in terms of measure theoretical concepts, i.e. in terms of σ-fields and σ-additive measures on σ-fields, and not probabilistic if more is needed to formulate them.

Consider now the Borel-Paradox situation in itself, as part of pure mathematics. Then the question is why one may think that the uniform length measure on a great circle is determined probabilistically by the surface measure on the sphere. One reason, we claim, is that when one thinks about the relation of the length measure and the surface measure, one might not distinguish carefully between the length measure being determined *probabilistically* (via conditionalization) and being determined by *some* mathematical condition. By the length measure "being determined probabilistically" we mean it being deducible from the surface measure referring only to measure theoretic concepts. Thus we may think *correctly* that the uniform distribution on the meridian and the surface measure are related in a very tight, natural way, but we might not realize that the link is not probabilistic.

This happens for instance when one "feels intuitively" that the rotational symmetry of the Borel–Kolmogorov situation singles out the uniform probability on a great circle as the only one that "matches" the uniform measure on the sphere (Myrvold 2014, Sect. 3.2). This feeling is justified in the sense that it can be translated into precise mathematical terms: The uniform measure on a great circle is singled out indeed as the (unique) measure that is invariant with respect to the natural "subgroup" of rotations (in the plane of the circle) of the full group of rotations in the three dimensional space, with respect to which the surface measure is invariant.[1] The important point to realize however is that this link between the surface measure and the measure on the circle is non-probabilistic, it cannot be stated in measure theoretic terms only: one needs the theory of (topological) groups to obtain the length measure this way. Thinking that the uniform length measure on a great circle is determined probabilistically by the surface measure on the sphere is therefore a fallacious intuition.

This fallacy can be made more explicit: Since probability theory is specified in the Kolmogorovian axiomatization as a probability measure space, a concept is probabilistic only if it is invariant with respect to isomorphisms of probability measure spaces. A probability measure space $(X, \mathcal{S}, p)$ is called a *standard probability space* if X is a complete, separable metric space and $\mathcal{S}$ is the completion of the Borel σ-algebra of X. Standard, non-atomic probability spaces are isomorphic to the unit interval with the Lebesgue measure on it (Aaronson 1997, Chap. 1, p. 3). Since the sphere with its uniform surface measure is a standard, measure theoretically non-atomic probability space, it is isomorphic as a probability measure space to the unit interval with the Lebesgue measure on it. Under this isomorphism the problem of what the conditional probability distribution on the great circle is, gets translated faithfully (i.e. without any loss or distortion of its probabilistic content), into a problem about the conditional prob-

[1] There are some mathematical subtleties involved in how this can be done; see Appendix 3 in the Appendix section.

ability distribution on a probability zero set in the unit interval. Since in this probabilistically fully equivalent translation of the problem we might not have available any symmetry that we could rely on to specify a measure on that probability zero set, it becomes clear that the rotational symmetry of the conditional probability on the great circle is not a feature of the conditional probability that can be regarded as determined probabilistically by the assumption that the distribution on the sphere is the uniform one.

Another tight link can be established between the uniform measure on the sphere and the uniform length measure on a great circle if we think of the sphere, of the great circle and of their relation not group theoretically but geometrically: regarding a great circle as a closed one-dimensional differentiable submanifold of the sphere viewed as a two dimensional differentiable manifold embedded in three dimension, the uniform measure on both the great circle and the sphere can be obtained from the Lebesgue measure in three dimension in a canonical manner via standard techniques in differential geometry (Morvan 2008, Sects. 3.1, 5.3–5.4]. Again, this link between the uniform measures on the sphere and on a great circle is very natural but cannot be regarded as *probabilistic* because concepts of differential geometry are crucial and indispensable in it and these concepts are not purely measure theoretic.

Distinguishing between probabilistic and non-probabilistic in terms of measure theoretic isomorphism invariance helps to clarify further the status of the particular versions of the respective conditional expectations that yield the uniform and non-uniform probability distributions on a single great circle: A measure theoretic isomorphism between probability spaces sets up a one-to-ne correspondence between versions of conditional expectations in the two probability spaces that are determined by σ-fields that are isomorphic under the measure theoretic isomorphism. The conditional probabilities of events given by versions that are related in this way are the same—this is in harmony with defusing the Description Relativity Worry. But the product structure of the probability measure space formed by the sphere with its surface measure is not invariant with respect to measure theoretic isomorphisms. Hence "singling out" the particular versions that yield the conditional probabilities on a single great circle depends on the specific, not purely measure theoretic properties of the sphere with its surface measure, and the versions are therefore not determined purely probabilistically by the prior uniform probability on the sphere and the conditioning σ-field.

7.2 Probability theory as mathematical model

Like other mathematical theories, probability theory also can be used to describe phenomena. In such applications of probability theory, the random events A in $\mathcal{S}$ are related to other entities, and the truth conditions of the statement $p(A) = r$ have to be specified. In an application, probability theory thus becomes a mathematical *model* of a certain phenomenon. The phenomenon itself can be either mathematical or non-mathematical. A specific probability space is a good model of a phenomenon if the statements $p(A) = r$ are indeed true (in the sense of the specified condition that is part of the model).

When one looks at the Borel–Kolmogorov Paradox from the perspective of the concept of application so described, one has to ask what the sphere with the uniform

distribution on it is a probabilistic model of; i.e. what the phenomenon is that the probabilistic model describes and how precisely the mathematical theory is related to the phenomenon in question. There are several conceivable scenarios here. Somewhat surprisingly, the papers discussing the Borel–Kolmogorov Paradox typically do not specify any.[2] This is unfortunate because without knowing what precisely the probability space is a model of, it is impossible to assess whether certain intuitions about the probabilistic model are correct or not.

A possible scenario, which is probably closest to how the Borel–Kolmogorov situation is tacitly interpreted in the literature, is the following. It is assumed that a specific mathematical algorithm yields points on the surface of the two dimensional unit sphere in the three dimensional Euclidean space. The uniform probability measure on the sphere can then be thought of as a model of generating points on the sphere in the sense of relative frequencies: Running the algorithm N times one can compute the number $r(A, N)$ of the generated points falling in a measurable set A of the sphere, and one can also compute the limit of the ratio $\frac{r(A,N)}{N}$ as $N \to \infty$. If the limit exists and is proportional to the measure of the set A in the surface measure for any measurable set A, then the sphere with the uniform surface measure is a good probabilistic model of the point generating algorithm. Note that since generating points on the two dimensional sphere (more generally: on the N-dimensional sphere) with uniform distribution is important in Monte Carlo simulations run on computers, the problem of which algorithms produce such points has been studied extensively and several such algorithms have been found (Muller 1959; Sibuya 1964; Marsaglia 1972; Tashiro 1977; see also Feller 1966, pp. 29–33).

Viewed from the perspective of this application, both the $(\mathcal{O}, q_{\mathcal{O}})$-conditional and the $(\mathcal{M}, q_{\mathcal{M}})$-conditional distributions on great circles are intuitively correct: Given the concept of conditional probability defined by conditional expectation (Sect. 2.4), the full claim about the $(\mathcal{O}, q_{\mathcal{O}})$-conditional probabilities is: Given the σ-field $\mathcal{O}$, the (conditional) probability measure on *all* circles in $\mathcal{O}$ (which are all parallel to a great circle C on the sphere) is the uniform probability on the circles (Remark 5). Since conditional probabilities are required to satisfy the theorem of total probability by definition, the major content of this claim is that if we "add up" the conditional probabilities on these circles, then we obtain the uniform measure on the surface. Thus, if we wish to generate points on the surface of the sphere using some mathematical algorithm, then if the points are generated in such a way that their distribution is uniform on all circles in $\mathcal{O}$ (i.e. uniform in the longitude variable ϕ) then the distribution of the generated points will be uniform on the surface of the sphere. This is intuitively correct: the circles are all parallel, they are all disjoint, their lengths contribute equally to the surface measure.

The full claim about the $(\mathcal{M}, q_{\mathcal{M}})$-conditional probabilities is: Given the σ-field $\mathcal{M}$ generated by (measurable sets of) all meridian circles (which all share the same *North* and *South* poles), the conditional probability measure on *all* meridian circles in $\mathcal{M}$ is the probability measure given by the density function $\cos\theta$ on the meridian circles, where θ is the latitude variable on every meridian circle (Remark 5). Just as in

[2] Rescorla's paper Rescorla (2014) being an exception, see the end of this section.

the case of $(\mathcal{O}, q_{\mathcal{O}})$-conditional probabilities, the major content of this claim is that if we "add up" the conditional probabilities on these meridian circles, then we obtain the uniform measure on the surface. Thus, suppose points are generated on the surface of the sphere by some mathematical algorithm. If the points are generated in such a way that their distribution is given by the density $\cos\theta$ on all meridian circles in $\mathcal{M}$, then the distribution of the generated points will be uniform on the surface of the sphere. To formulate this negatively: If we choose a generating algorithm that produces points on all meridian circles according to the uniform distribution in the latitude variable θ, then the points will *not* be uniformly distributed on the sphere—simply because the meridian circles are positioned on the sphere in a very specific way: unlike the parallel, disjoint circles in $\mathcal{O}$, the meridian circles are not disjoint: they all have the same *North* and *South Poles* in common, which entails that they are "crammed" around the poles. Hence, if the points are generated uniformly in θ on all meridian circles, then more points are generated closer to the poles, and thus the points generated this way will accumulate more around the poles. This makes perfect intuitive sense and can in fact be illustrated by computer simulation (Weisstein 2015).

That the $\mathcal{A}$-conditional expectation and hence the $(\mathcal{A}, q_{\mathcal{A}})$-conditional probability distribution on a *single*, *fixed* great circle C is undetermined if $\mathcal{A}$ is the four element Boolean algebra consisting of C, its complement, the empty set and the whole sphere, also is correct intuitively: the sphere with its uniform surface measure can be a perfectly good probabilistic model of the behavior of the mathematical algorithm in the sense described, together with *any* behavior of the algorithm on the *fixed*, *single* great circle C—precisely because C is measure zero in the surface measure. In other words, saying that the surface measure is a good probabilistic model of the algorithm simply does not contain enough information to infer *probabilistically* that the algorithm behaves in any particular way on the specified great circle C. Furthermore, it is clear that this behavior cannot be found out via any a priori reasoning (such as some form of Principle of Indifference for instance). One just would have to look at the specific algorithm generating the points on the sphere and see what distribution it generates on *that particular* C. Whatever this distribution is, it can however be accommodated as conditional probability using the theory of conditional expectations by taking the appropriate version of the $\mathcal{A}$-conditional expectation.

One also can envisage a scenario in which the sphere with its uniform measure is the model not of objective frequencies but of subjective degrees of beliefs (credences). Rescorla's recent paper Rescorla (2014) analyzes the Borel–Kolmogorov Paradox from this perspective, and Easwaran's defusing of the paradox also is under the subjective interpretation Easwaran (2008). Rescorla's main claim is that "The Borel–Kolmogorov paradox is a Fregean 'paradox of identity'." (Rescorla 2014, p. 16) In harmony with this, Rescorla embraces (Rescorla 2014, p. 14) the conditional probabilities' sensitive dependence on the conditioning σ-field as a sign of subjective degrees of beliefs' dependence on how the events are represented to the agent. In his view this dependence is not irrational because

> Credences are rationally sensitive to the way that one represents an event, so it is not surprising that different conditional probabilities arise when one represents the conditioning event differently. (Rescorla 2014, p. 16)

An assessment of Rescorla's proposal would require going into the details of the Fregean paradox of identity, together with an explication of the nature of credences and with an elaboration of the relation of credences to probability theory, which we cannot undertake here. But we agree with Rescorla's general methodological stand: that the analysis of the Borel–Kolmogorov Paradox should take into account how one interprets probability (Rescorla 2014, p. 14) (in our terminology: what the application of probability theory is). The application described in this section is under the objective, frequency interpretation, which Rescorla leaves for others to analyze (Rescorla 2014, p. 14). We have seen that under this interpretation of probability both the $(\mathcal{O}, q_{\mathcal{O}})$-conditional and the $(\mathcal{M}, q_{\mathcal{M}})$-conditional distributions on the great circle are intuitively correct. Thus, although for reasons different from his, we also agree with Rescorla that "... the Borel–Kolmogorov paradox is not remotely paradoxical." (Rescorla 2014, p. 17)

8 Closing comments

The Borel–Kolmogorov Paradox was formulated by Borel in 1909, at a time, when the conceptual foundations of probability theory were not yet entirely clear. In particular, the notion of conditional probability was then restricted to the Bayes' rule, which is a very special and limited concept, not revealing the real conceptual structure of conditionalization. It was only as a result of Kolmogorov's work that the abstract conceptual structure of probability theory and of conditionalization became clarified. Kolmogorov's work was the result of a long development (see Doob 1996 for the major conceptual steps in the history of rigorous probability theory), and its significance is not just that it is the natural end of a long development but that it is the beginning of a new phase of probability theory: Kolmogorov's work established a link between probability theory and functional analysis. This opened the way for developing probabilistic concepts that are indispensable in probabilistic modeling of certain phenomena (martingales, stochastic processes). The notion of conditional expectation is crucial in this further development in probability theory.

The starting point of our analysis of the Borel–Kolmogorov Paradox in this paper was adopting the Kolmogorovian view, which is standard in today's probability theory: that the suitable technical device of conditionalization is the concept of conditional expectation. The main goal of the theory of conditional expectations

> ... is the systematic development of a notion of conditional probability that covers conditioning with respect to events of probability 0. This is accomplished by conditioning with respect to collections of events—that is, with respect to σ-fields. (Billingsley 1995, p. 432).

The concept of conditional expectation with respect to a σ-field was developed by Kolmogorov (1933) soon after the main tool for it, the Radon–Nikodym theorem, had been found in 1930, and Kolmogorov's resolution of the Borel Paradox is also based on the idea of emphasizing conditioning with respect to σ-fields:

> [The Borel Paradox] shows that the concept of a conditional probability with regard to an isolated given hypothesis whose probability equals 0 is inadmissible. For we can obtain a probability distribution [...] on the meridian circle only if we regard this circle as an element of a decomposition of the entire spherical surface into meridian circles with the given poles. (Kolmogorov 1933, p. 51)

Kolmogorov's wording of his resolution of the Borel–Kolmogorov Paradox is slightly misleading however because it makes the impression that only by taking the σ-field $\mathcal{M}$ generated by (measurable sets of) meridian circles and calculating the conditional probabilities using the conditional expectation this σ-field defines (obtaining this way the non-uniform distribution on meridian circles) will yield a conditional probability on a great circle that is intuitively correct. We have seen however that one also can take the σ-field $\mathcal{O}$ generated by (measurable sets of) circles parallel to a given great circle and compute the corresponding conditional probabilities. Doing this, one obtains different but intuitively not less correct conditional probabilities. It does not make sense to ask "Which one of the σ-fields $\mathcal{M}$ and $\mathcal{O}$ define the 'correct' conditional probabilities?" The algebras $\mathcal{M}$ and $\mathcal{O}$ represent different conditioning circumstances and the conditional probabilities they lead to are answers to different questions—not different answers to the same question. In certain applications $\mathcal{M}$, in certain other applications $\mathcal{O}$ might represent some circumstances that are described correctly by the corresponding conditional probabilities. This is an advantage, showing the flexibility of probability theory in modeling phenomena. Thus worries about the dependence of conditional probabilities on the conditioning algebras seem to us to be misguided. There is no "absolute" notion of conditional probability—conditional probabilities are truly *conditional*: they depend on a full set of conditions, i.e. on a σ-field. This is so also when one uses Bayes' rule to calculate conditional probabilities; in this specific case the dependence of the conditional probability on the full four element Boolean subalgebra generated by the single conditioning random event featuring in Bayes' rule is just not quite transparent.

Thus, under close and careful scrutiny, the "paradox" in the Borel–Kolmogorov Paradox evaporates: There is no clash between the *correct* intuition about what the conditional probabilities with respect to probability zero events are and the technically proper concept of conditionalization via conditional expectation.

Acknowledgments We thank three anonymous referees for their helpful suggestions to improve the first version of the manuscript. Research supported by the National Research, Development and Innovation Office, K 115593 and K 100715. M. Rédei thanks the Institute of Philosophy of the Hungarian Academy of Sciences, with which he was affiliated as Honorary Research Fellow while this paper was written.

Appendix

Appendix 1: Computing the $(\mathcal{O}, q_{\mathcal{O}})$-conditional probability on a great circle

Keeping the notations (in particular the polar coordinates) introduced in Sect. 4, we verify that that $\mathcal{E}(\cdot \mid \mathcal{O})$ given by (72) is a version of the $\mathcal{O}$-conditional expectation.

Note first that, according to standard integral transformation theorems, for all integrable functions f we have

$$\int_S f d\mu = \int_0^{\pi} \int_0^{2\pi} f(\phi, \theta) \cdot \left| \frac{\partial(x, y, z)}{\partial(\phi, \theta, r)} \right| d\phi d\theta \tag{92}$$

where μ is the surface measure on the sphere S and $\left| \frac{\partial(x,y,z)}{\partial(\phi,\theta,r)} \right|$ is the absolute value of the Jacobian determinant

$$\begin{aligned} \left| \frac{\partial(x, y, z)}{\partial(\phi, \theta, r)} \right| &= \begin{vmatrix} \cos\phi \sin\theta & -r \sin\phi \sin\theta & r \cos\phi \sin\theta \\ \sin\phi \sin\theta & r \cos\phi \sin\theta & r \sin\phi \cos\theta \\ \cos\theta & 0 & -r \sin\theta \end{vmatrix} \\ &= r^2 \sin\theta \quad (\text{for } 0 \leq \theta \leq \pi) \end{aligned} \tag{93}$$

(Note that if $0 \leq \theta \leq 2\pi$, then $\left| \frac{\partial(x,y,z)}{\partial(\phi,\theta,r)} \right| = r^2 |\sin\theta|$). The volume of the unit sphere is 2π thus after normalizing the surface measure we get

$$\int_S f d\mu = \int_S f \frac{d\nu}{2\pi} = \frac{1}{2\pi} \int_0^{\pi} \int_0^{2\pi} f(\phi, \theta) \cdot \sin\theta \, d\phi d\theta \tag{94}$$

We have to verify that, for $Z = [0, 2\pi] \times A$ in $\mathcal{O}$ we have

$$\int_Z \mathcal{E}(f \mid \mathcal{O}) \, d\mu = \int_Z f \, d\mu \tag{95}$$

The calculation:

$$\int_Z \mathcal{E}(f \mid \mathcal{O}) \, d\mu = \frac{1}{2\pi} \int_A \int_0^{2\pi} \mathcal{E}(f \mid \mathcal{O}) \sin\theta \, d\phi d\theta \tag{96}$$

$$= \frac{1}{2\pi} \int_A \int_0^{2\pi} \frac{1}{2\pi} \int_0^{2\pi} f(\phi, \theta) \, d\phi \sin\theta \, d\phi d\theta \tag{97}$$

$$= \frac{1}{2\pi} \int_A \frac{\sin\theta}{2\pi} \int_0^{2\pi} \int_0^{2\pi} f(\phi, \theta) \, d\phi \, d\phi d\theta \tag{98}$$

$$= \frac{1}{2\pi} \int_A \sin\theta \int_0^{2\pi} f(\phi, \theta) \, d\phi d\theta \tag{99}$$

$$= \frac{1}{2\pi} \int_A \int_0^{2\pi} f(\phi, \theta) \cdot \sin\theta \; d\phi d\theta \tag{100}$$

$$= \int_Z f \; d\mu \tag{101}$$

Appendix 2: Computing the $(\mathcal{M}, q_{\mathcal{M}})$-conditional probability on a great circle

Again, we have to verify that $\mathcal{E}(\cdot \mid \mathcal{M})$ given by (81) is a version of the $\mathcal{M}$-conditional expectation. This requires to show that for $Z = A \times [0, \pi] \in \mathcal{M}$ (where A is symmetric) we have

$$\int_Z \mathcal{E}(f \mid \mathcal{M}) \; d\mu = \int_Z f \; d\mu \tag{102}$$

The calculation:

$$\int_Z \mathcal{E}(f \mid \mathcal{M}) \; d\mu = \frac{1}{2\pi} \int_0^{\pi} \int_A \mathcal{E}(f \mid \mathcal{M}) \sin\theta \; d\phi d\theta \tag{103}$$

$$= \frac{1}{2\pi} \int_0^{\pi} \int_A \frac{1}{2} \int_0^{2\pi} f(\phi, \theta) |\sin\theta| \; d\theta \sin\theta \; d\phi d\theta \tag{104}$$

$$= \frac{1}{2\pi} \int_0^{\pi} \frac{\sin\theta}{2} \int_A \int_0^{2\pi} f(\phi, \theta) |\sin\theta| \; d\theta \; d\phi d\theta \tag{105}$$

$$= \frac{1}{2\pi} \int_0^{\pi} \frac{\sin\theta}{2} \cdot 2 \int_A \int_0^{\pi} f(\phi, \theta) \sin\theta \; d\theta \; d\phi d\theta \tag{106}$$

$$= \frac{1}{2\pi} \int_0^{\pi} \sin\theta \; d\theta \cdot \int_A \int_0^{\pi} f(\phi, \theta) \sin\theta \; d\theta \; d\phi \tag{107}$$

$$= \frac{1}{2\pi} \int_A \int_0^{\pi} f(\phi, \theta) \sin\theta \; d\theta \; d\phi \tag{108}$$

$$= \frac{1}{2\pi} \int_0^{\pi} \int_A f(\phi, \theta) \sin\theta \; d\phi \; d\theta \tag{109}$$

$$= \int_Z f \; d\mu \tag{110}$$

Remark 6 If we define the σ-algebra $\mathcal{M}'$ as the set of all meridian and half-meridian circles:

$$\mathcal{M}' = \left\{ A \times [0, \pi] : \; A \subseteq [0, 2\pi] \text{ is measurable} \right\} \tag{111}$$

(with A not necessarily symmetric), then the $\mathcal{M}'$-conditional expectation is given by

$$\mathcal{E}(f \mid \mathcal{M}')(\phi, \theta) = \int_0^{\pi} f(\phi, \theta) \sin\theta \; d\theta \tag{112}$$

This is verified by the following equations:

$$\int_Z \mathcal{E}(f \mid \mathcal{M}')\, d\mu = \frac{1}{2\pi} \int_0^\pi \int_A \mathcal{E}(f \mid \mathcal{M}') \sin\theta \; d\phi d\theta \tag{113}$$

$$= \frac{1}{2\pi} \int_0^\pi \int_A \int_0^\pi f(\phi, \theta) \sin\theta \; d\theta \sin\theta \; d\phi d\theta \tag{114}$$

$$= \frac{1}{2\pi} \int_0^\pi \sin\theta \; d\theta \cdot \int_A \int_0^\pi f(\phi, \theta) \sin\theta \; d\theta \; d\phi \tag{115}$$

$$= \frac{1}{2\pi} \int_A \int_0^\pi f(\phi, \theta) \sin\theta \; d\theta \; d\phi \tag{116}$$

$$= \frac{1}{2\pi} \int_0^\pi \int_A f(\phi, \theta) \sin\theta \; d\phi \; d\theta \tag{117}$$

$$= \int_Z f \; d\mu \tag{118}$$

Appendix 3: The group-theoretic relation of the sphere and of a great circle on the sphere

The two-dimensional sphere possesses a symmetry represented by the topological group $SO(3)$ of rotations in the three dimensional Euclidean space; however, the sphere itself *cannot* be considered as a topological group [see Megía (2007) and references there]. If by "uniform measure on the sphere" we mean the σ-additive *Lebesgue* measure, then this can be defined as the *unique* measure on the sphere which is invariant with respect to the action of $SO(3)$ on the sphere. While this seems intuitively obvious, it is in fact a non-trivial mathematical theorem having a non-trivial proof; furthermore, it is an *open problem* whether the σ-additive uniform *Borel* measure on the sphere also is the unique measure invariant with respect to $SO(3)$ (Kharazishvili 1997).

Any great circle also possesses a symmetry represented by the group $SO(2)$ of rotations in the two dimensional plane of the circle; in fact the circle *can* be identified with the group $SO(2)$ itself. The intuitive argument showing that the sphere itself cannot be viewed as the topological group SO(3) whereas the circle can be identified with SO(2) is that if we designate some special point of the circle, then every rotation of the circle is uniquely identified by the point that this special point is sent to, while if we designate some special point of the sphere, then a rotation of the sphere is not uniquely identified by the point that this special point is sent to, because we can rotate around the axis through this special point as well. $SO(2)$ is a compact topological group and can be thought of as a "subgroup" of $SO(3)$, although, strictly speaking $SO(2)$ is not a subgroup of $SO(3)$: the group $SO(3)$ is the set of 3×3 matrices having unit determinant, $SO(2)$ is the set of 2×2 matrices having unit determinant, thus these groups have different unit elements e_3 and e_2, respectively. But $SO(2)$ can be embedded into $SO(3)$ by an injective group homomorphism h; elements of the form $e_3 \cdot h(g)$ in $SO(3)$, with $g \in SO(2)$ form a subgroup. Since on every compact topological group there exists a *unique* group invariant normalized measure, the so-called Haar

measure, there exists a unique rotational invariant measure on $SO(2)$—this is the length measure on the circle. Standard references for the Haar measure are Nachbin (1965) and Halmos (1950, Chap. XI), for a more recent presentation see Deitmar and Echterhoff (2009)).

References

Aaronson, J. (1997). *An introduction to infinite ergodic theory*. Mathematical surveys and monographs (Vol. 50). Rhode Island: American Mathematical Society.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York: Wiley.

Bogachev, V. I. (2007). *Measure theory* (Vol. II). Berlin: Springer.

Borel, E. (1909). *Éleménts de la Théorie des Probabilités*. Paris: Librairie Scientifique A. Herman & Fils. English translation: by J. Freund, "Elements of the Theory of Probability", Englewood Cliffs, 1956, Prentice-Hall.

de Finetti, B. (1972). *Probability, induction, and statistics*. New York: Wiley.

Deitmar, A., & Echterhoff, S. (2009). *Principles of harmonic analysis*. New York: Universitext. Springer.

Doob, J. (1996). The development of rigor in mathematical probability theory (1900-1950). *American Mathematical Monthly*, *103*(7), 586–595.

Easwaran, K. (2008). The foundations of conditional probability. PhD thesis, University of California at Berkeley.

Feller, W. (1966). *An introduction to probability theory and its applications* (Vol. 2). New York: Wiley, 2nd edition: 1971. First edition: 1966.

Gyenis, Z., & Rédei, M. (2015a). Defusing Bertrand's Paradox. *The British Journal for the Philosophy of Science*, *66*, 349–373.

Gyenis, Z., & Rédei, M. (2015b). Why Bertrand's Paradox is not paradoxical but is felt so. In U. Maki, S. Ruphy, G. Schurz, & I. Votsis (Eds.), *Recent developments in the philosophy of science: EPSA13* (pp. 265–276). Helsinki: Springer.

Gyenis, Z., & Rédei, M. (2016). General properties of general Bayesian learning. *Erkenntnis*. submitted.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, *137*, 273–333.

Halmos, P. (1950). *Measure theory*. New York: D. Van Nostrand.

Harper, W. (1975). Rational belief change, Popper functions, and counterfactuals. *Synthese*, *30*, 221–262.

Howson, C. (2014). Finite additivity, another lottery paradox, and conditionalization. *Synthese*, *191*, 989–1012.

Jaynes, E. T. (2003). *Probability theory. The logic of science* (G. Larry Bretthorst, Ed.). Cambridge: Cambridge University Press.

Jeffrey, R. C. (1965). *The logic of decision* (1st ed.). Chicago: The University of Chicago Press.

Kharazishvili, A. B. (1997). On the uniqueness of Lebesgue and Borel measures. *Journal of Applied Analysis*, *3*, 49–66.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. English translation: Foundations of the Theory of Probability, (Chelsea, New York, 1956).

Loéve, M. (1963). *Probability theory* (3rd ed.). Princeton: D. Van Nostrand.

Makinson, D. (2011). Conditional probability in the light of qualitative belief change. *Journal of Philosophical Logic*, *40*, 121–153.

Marchand, J.-P. (1981) Statistical inference in quantum mechanics. In K.E. Gustafson, W. P. Reinhard (Eds.), *Quantum Mechanics in Mathematics, Chemistry, and Physics* (pp. 73–81). New York: Plenum Press. Proceedings of a special session in mathematical physics organized as a part of the 774th meeting of the American Mathematical Society, held March 27–29, 1980, in Boulder, CO.

Marchand, J.-P. (1977). Relative coarse-graining. *Foundations of Physics*, *7*, 35–49.

Marchand, J.-P. (1982). Statistical inference in non-commutative probability. *Rendiconti del Seminario Matematico e Fisico di Milano*, *52*, 551–556.

Marsaglia, G. (1972). Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, *43*, 645–646.

Megía, I. S.-M. (2007). Which spheres admit a topological group structure? *Revista de la Real Academia de Ciencias de Zaragoza*, *62*, 75–79.

Morvan, J.-M. (2008). *Generalized curvatures*. Springer series in geometry and computing. Berlin: Springer.

Muller, M. E. (1959). A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the Association for Computing Machinery*, *2*, 19–20.
Myrvold, W. (2014). You can't always get what you want: Some considerations regarding conditional probabilities. *Erkenntnis*. Forthcoming, online August 5, 2014, doi:10.1007/s10670-014-9656-3.
Nachbin, L. (1965). *The Haar integral*. Princeton, NJ: D. Van Nostrand.
Popper, K. (1995). *The logic of scientific discovery*. London: Routledge. First published in English in 1959 by Hutchinson Education.
Popper, K. (1938). A set of independent axioms for probability. *Mind*, *47*, 275–277.
Popper, K. (1955). Two autonomous axiom systems for the calculus of probabilities. *The British Journal for the Philosophy of Science*, *6*, 51–57.
Proschan, M. A., & Presnell, B. (1998). Expect the unexpected from conditional expectation. *The American Statistician*, *52*(3), 248–252.
Rao, M. M. (2005). *Conditional measures and applications*, 2nd revised and expanded edition. Boca Raton: Chapman & Hall/CRC.
Rao, M. M. (1988). Paradoxes in conditional probability. *Journal of Multivariate Analysis*, *27*, 434–446.
Rényi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiae Hungaricae*, *6*, 268–335.
Rescorla, M. (2014). Some epistemological ramifications of the Borel-Kolmogorov Paradox. *Synthese*. Forthcoming. Published online November 20, 2014. doi:10.1007/s11229-014-0586-z.
Rosenthal, J. S. (2006). *A first look at rigorous probability theory*. Singapore: World Scientific.
Seidenfeld, T. (2001). Remarks on the theory of conditional probability: Some issues of finite versus countable additivity. In V. F. Hendricks (Ed.), *Probability theory* (pp. 167–178). Dordrecht: Kluwer Academic Publishers.
Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2001). Improper regular conditional distributions. *The Annals of Probability*, *29*, 1612–1624.
Sibuya, M. (1964). A method for generating uniformly distributed points on n-dimensional spheres. *Annals of the Institute of Statistical Mathematics*, *14*, 81–85.
Tashiro, Y. (1977). On methods for generating uniform random points on the surface of a sphere. *Annals of the Institute of Statistical Mathematics*, *29*, 295–300.
van Fraassen, B. C. (1976). Representation of conditional probabilities. *Journal of Philosophical Logic*, *5*, 417–430.
Weisstein, E. (2015). Wolfram MathWorld. A free resource from Wolfram Research built with Mathematica technology: http://mathworld.wolfram.com/. Entry: Sphere Point Picking.