Consistent reasoning about a continuum of hypotheses on the basis of finite evidence

Jochen Rau*
Rhönstraße 70, 60385 Frankfurt, Germany
(Dated: February 1, 2008)

In the modern Bayesian view classical probability theory is simply an extension of conventional logic, i.e., a quantitative tool that allows for consistent reasoning in the presence of uncertainty. Classical theory presupposes, however, that—at least in principle—the amount of evidence that an experimenter can accumulate always matches the size of the hypothesis space. I investigate how the framework for consistent reasoning must be modified in non-classical situations where hypotheses form a continuum, yet the maximum evidence accessible through experiment is not allowed to exceed some finite upper bound. Invoking basic consistency requirements pertaining to the preparation and composition of systems, as well as to the continuity of probabilities, I show that the modified theory must have an internal symmetry isomorphic to the unitary group. It thus appears that the only consistent algorithm for plausible reasoning about a continuum of hypotheses on the basis of finite evidence is furnished by quantum theory in complex Hilbert space.

PACS numbers: 02.50.Cw, 03.65.Ta, 03.67.-a

I. INTRODUCTION

In the modern Bayesian view classical probability theory with its two key ingredients (i) Bayes' learning rule and (ii) maximum entropy priors, is nothing but an extension of conventional logic, i.e., a quantitative tool that allows for consistent reasoning also in the presence of uncertainty [1, 2]. Probabilities are no longer defined as limits of relative frequencies but as "degrees of belief" that are subject to certain consistency requirements [3], and that can be legitimately assigned not just to ensembles but also to individual systems. Bayesian probability theory is thus more broadly applicable than the orthodox frequentist approach, while yielding identical results in those cases where a large N limit exists.

Quantum theory is inherently probabilistic: Does it therefore, too, lend itself to a Bayesian interpretation [4, 5]? More specifically, does quantum theory, too, represent some kind of "optimal algorithm" for plausible reasoning in a certain, yet to be specified setting? There are indications that this might be the case, as quantum theory has been linked to concepts such as a modified propositional calculus, "learning", and—most recently—information processing: (i) One of the earliest attempts (long before the advent of modern Bayesianism) to axiomatize quantum theory started from a generalisation of classical propositional calculus, relaxing the requirement that all propositions be jointly decidable and resulting in a mathematical structure dubbed "quantum logic" [6, 7, 8, 9]; the key result of this approach being that propositions within (an irreducible building block of) such a "quantum logic" can always be identified with subspaces of a Hilbert space over some skew field [10]. (This approach fails, however, to give a compelling argument why the skew field should be the complex numbers, and does not work for dimension two.) (ii) The discontinuous change of the density matrix upon quantum measurement has been shown to be closely related to Bayes' learning rule [11]. (iii) Ongoing research in the fast-growing field of quantum information and quantum computation keeps revealing intimate connections of quantum theory with, and its potential power for, information processing [12, 13, 14, 15]. There is even a recent proposal to reduce the key features of quantum theory—albeit not the full Hilbert space structure—to a small number of purely information-theoretic constraints [16].

In this paper I attempt to pinpoint the circumstances under which, and the sense in which, the basic laws of quantum theory may indeed be considered an "optimal" set of rules to conduct plausible reasoning in the presence of uncertainty. But how is this possible if classical Bayesian theory is already thought to be *the* universal algorithm? The basic idea is the following. In a probabilistic model every proposition can be built up, through logical operations, from a certain minimal set—called the "hypothesis space"—of elementary propositions. Classical probability theory assumes that all these elementary propositions are jointly decidable: An experiment can be devised (at least in principle) by which the truth values of all elementary propositions can be jointly ascertained. Arbitrary repetitions of this experiment will reproduce with certainty the same result. Such a most refined experiment yields as evidence a string of truth values 0 or 1. The length of this string is a measure for the maximum amount of evidence that can be garnered from experiment. It equals the cardinality of the hypothesis space. At least in theory, therefore, the amount of evidence that an experimenter can accumulate matches the size of the hypothesis space.

In quantum theory the situation is radically different [17]. There are propositions pertaining to non-commuting observables that are not jointly decidable.

*Electronic address: jochen.rau@web.de

For a finite-dimensional quantum system the total amount of reproducible evidence that can be garnered from experiment is bounded from above by the Hilbert space dimension and hence by a finite number; whereas the hypothesis space comprises all possible pure states and hence is a continuous manifold. The amount of evidence that any experimenter can accumulate is thus strictly *smaller* than the hypothesis space, not due to practical limitations but as a matter of principle; maximal information is not complete [18, 19].

It is the aim of this paper to show that not only does quantum theory imply a mismatch between hypothesis space and available evidence, but the converse is also true: Whenever one is confronted with a situation where hypotheses form a continuum but total evidence is not allowed, as a matter of principle, to exceed a finite upper bound then plausible reasoning about that continuum of hypotheses, if it is to satisfy some basic consistency requirements, must necessarily follow the rules of quantum theory.

The proof of this claim presupposes that some basic consistency requirements for plausible reasoning—with the notable exception of "joint decidability"—carry over from the classical case; they will be detailed below. As the principal subject of inquiry I will then introduce the group of "consistency-preserving" transformations in the continuous hypothesis space. Analysis of this group, which to a good part amounts to a simple dimension-counting exercise, reveals that it must be isomorphic to the unitary group $U(d)$ where d is the finite upper bound on the evidence. This mandates the use of complex Hilbert space, and the identification of propositions with its subspaces, as the sole consistent framework for plausible reasoning.

Inferring the Hilbert space structure of quantum theory by means of a dimensional analysis has been proposed before [20]. Like the proof given below, that earlier proposal invoked the correspondence between probability distributions and measurements (state preparation), rules for the composition of systems, and the continuity of probabilities; it focused on demonstrating that the manifold of (non-normalised) states has dimension $P(d) = d^2$. However, it provided a rigorous proof only of $P(d) = d^\mu$, $\mu \in \mathbf{N}$, the cases $\mu \geq 3$ being excluded merely on the basis of a non-rigorous, albeit plausible, "simplicity" requirement. In contrast to the approach presented here, the earlier proposal did not include a systematic study of the structure group and its dimension. And finally, it made extensive use of the concepts "pure state" and "fiducial state", as well as of the language of linear vector spaces: notions that are inspired by quantum theory and already very suggestive of the structure to be derived, and that we will be trying to avoid here.

The remainder of this paper is organised as follows. In Section II we introduce the basic notions of hypotheses, probabilities, filters and transformations. The latter constitute a group, which will be the principal subject of our inquiry. Section III provides a precise definition of "maximum available evidence", and argues that it alone determines the appropriate mathematical framework for plausible reasoning; "evidence" is the sole parameter of the theory. Section IV constitutes the core of our analysis. We inspect carefully, and formulate a number of consistency requirements pertaining to, the preparation and composition of systems, as well as the continuity of probabilities. Thorough dimensional analysis then yields severe constraints on the structure group and leaves $U(d)$ as the only allowed choice. We also discuss how this result may change if any of our assumptions are relaxed. Finally, we wrap up our investigation with some concluding remarks in Section V. There is an appendix in which we give some technical proofs omitted in the main text.

II. BASIC NOTIONS

A. Hypotheses and probabilities

We are concerned with hypotheses about some given physical system. Some (but not all) of these hypotheses may be related by logical implication: $x \subseteq a$ means that if hypothesis x is true then hypothesis a is also true; hypothesis x is a "refinement" of hypothesis a. We shall denote the set of all possible refinements of a hypothesis a by

$$\mathcal{L}_a := \{x \mid x \subseteq a\} \ . \tag{1}$$

Logical implication constitutes a partial order: It is (i) reflexive, $x \subseteq x$; (ii) antisymmetric, $x \subseteq y,\ y \subseteq x \Rightarrow x = y$; and (iii) transitive, $x \subseteq y,\ y \subseteq z \Rightarrow x \subseteq z$. There is a unique null element $\emptyset$, sometimes called the "absurd hypothesis", which is always false and hence implies all others (ex falso quodlibet): $x = \emptyset \Leftrightarrow x \subseteq a\ \forall a$.

A probability distribution ρ assigns to each hypothesis a real number between 0 and 1. We shall denote the set of all probability distributions on $\mathcal{L}_a$ by $\mathcal{P}_a$. These two sets, $\mathcal{L}_a$ and $\mathcal{P}_a$, are dual to each other in the sense that any $\rho \in \mathcal{P}_a$ is completely specified by $\{\rho(x) \mid x \in \mathcal{L}_a\}$, and conversely any $x \in \mathcal{L}_a$ is completely specified by $\{\rho(x) \mid \rho \in \mathcal{P}_a\}$. In $\mathcal{P}_a$ there is a partial order mirroring that in $\mathcal{L}_a$, defined by

$$\rho \leq \sigma \ :\Leftrightarrow\ \rho(x) \leq \sigma(x)\ \forall x \in \mathcal{L}_a \ . \tag{2}$$

In keeping with the Bayesian spirit we do not make any reference to limits of relative frequencies but only demand that the assignment of probabilities satisfy a number of consistency requirements. Probability distributions must satisfy the common sense requirement that the more refined a hypothesis, the smaller its probability of being true; for any $x, y \in \mathcal{L}_a$,

$$x \subseteq y \ \Leftrightarrow\ \rho(x) \leq \rho(y)\ \forall \rho \in \mathcal{P}_a \ . \tag{3}$$

Probabilities are calibrated such that $\rho(\emptyset) = 0$; whereas they need not necessarily be normalised, i.e., the probability of the maximal element $a \in \mathcal{L}_a$ may be smaller than one.

When an observer assigns to a system either of the two probability distributions ρ or σ with respective "probability of probabilities" [5] $\mathrm{prob}(\rho)$ and $\mathrm{prob}(\sigma)$ then, on this meta-level, the resulting probability for a hypothesis x being true is given by the classical Bayes rule

$$\begin{aligned}\mathrm{prob}(x) &= \mathrm{prob}(x\,|\,\rho)\cdot\mathrm{prob}(\rho)+\mathrm{prob}(x\,|\,\sigma)\cdot\mathrm{prob}(\sigma)\\ &= \mathrm{prob}(x\mid \mathrm{prob}(\rho)\cdot\rho+\mathrm{prob}(\sigma)\cdot\sigma)\ ,\end{aligned}\tag{4}$$

where $\mathrm{prob}(x\,|\,\rho)=\rho(x)$ (and likewise for σ). Such mixing thus yields a new probability distribution which, being perfectly consistent, must also be contained in the set $\mathcal{P}_a$. The latter is therefore convex:

$$\rho,\sigma\in\mathcal{P}_a\ \Rightarrow\ t\rho+(1-t)\sigma\in\mathcal{P}_a\ \forall\ t\in[0,1]\ .\tag{5}$$

Since we do not require probability distributions to be normalised, arbitrary rescaling is permitted as long as probabilities never become greater than one:

$$\rho\in\mathcal{P}_a\ \Rightarrow\ s\rho\in\mathcal{P}_a\ \forall\ s\in[0,1/\rho(a)]\ .\tag{6}$$

B. Filters

There are further consistency requirements related to the processing of experimental evidence. We imagine an experiment—we call it a "filter"—that tests a certain hypothesis b and then keeps the system only if b is true, or else discards it if b is false. In the course of such an experiment the experimenter will acquire new information and consequently update the probability distribution in two steps, which are to be carefully distinguished: (i) upon learning that the filter has been applied, yet with outcome still unknown; and (ii) upon learning about the outcome. This can be summarized graphically as follows:

$$\rho\ \overset{\text{(i)}}{\rightarrow}\ \pi_b\rho\ \overset{\text{(ii)}}{\rightarrow}\ \begin{cases}\frac{1}{\rho(b)}\pi_b\rho & \text{if } b \text{ true}\\ 0 & \text{otherwise (system discarded)}\end{cases}\tag{7}$$

Step (i) introduces a—yet to be specified—map π_b whose required properties will be discussed below; whereas step (ii) is a simple rescaling that carries over directly from the classical Bayes rule.

After step (i) all post-filter (but pre-reading) probabilities are bounded from above by the system's survival probability,

$$\pi_b\rho\,(x)\le\rho(b)\ .\tag{8}$$

In general these equal their prior values only for the hypothesis being tested and its refinements,

$$\pi_b\rho\,(x)=\rho(x)\ \forall\rho\ \Leftrightarrow\ x\subseteq b\ .\tag{9}$$

Filtering must preserve the partial order of probability distributions,

$$\rho\le\sigma\ \Leftrightarrow\ \pi_b\rho\le\pi_b\sigma\ \forall b\ .\tag{10}$$

Finally, two filters can be applied in series (without intermediate reading of results). If one of the hypotheses being tested is a refinement of the other then one may just as well apply the finer filter only; the coarser filter becomes redundant:

$$b\subseteq a\ \Leftrightarrow\ \pi_b\circ\pi_a=\pi_a\circ\pi_b=\pi_b\ .\tag{11}$$

However, for arbitrary hypotheses not related by logical implication the order in which the respective filters are applied may matter, so we do *not* require that $\pi_b\circ\pi_a=\pi_a\circ\pi_b$ holds for every a,b.

Two hypotheses are said to "contradict" each other, $a\perp b$, if whenever one of them is true the other must be false. Operationally this means that successive application of the respective filters must always lead to the system being discarded:

$$a\perp b\ :\Leftrightarrow\ \pi_a\circ\pi_b=\pi_b\circ\pi_a=0\ .\tag{12}$$

A set of hypotheses $\{b_i\}$ shall be called a "set of alternative refinements" of a if they are mutually exclusive, $b_i\perp b_j\ \forall i\neq j$, while $b_i\subseteq a\ \forall i$; the set is "complete" if the refinements are also collectively exhaustive,

$$x\perp b_i\ \forall i\ \Leftrightarrow\ x\perp a\ .$$

An incomplete set of alternative refinements can always be made complete by adding to it the unique hypothesis "a, but not any of $\{b_i\}$". For a complete set of alternative refinements we require that the classical sum rule carry over,

$$\{b_i\}\prec a\ \Leftrightarrow\ \rho(a)=\sum_i\rho(b_i)\ \forall\rho\ ,\tag{13}$$

where we have defined "$\prec$" as meaning that $\{b_i\}$ is a complete set of alternative refinements of a.

If a system is described by a mixture of two probability distributions ρ,σ then application of the filter π_b leads to a posterior probability

$$\begin{aligned}\mathrm{prob}(x\,|\,\pi_b) &= \mathrm{prob}(x\,|\,\rho,\pi_b)\cdot\mathrm{prob}(\rho\,|\,\pi_b)+\\ &\quad+\mathrm{prob}(x\,|\,\sigma,\pi_b)\cdot\mathrm{prob}(\sigma\,|\,\pi_b)\end{aligned}\tag{14}$$

for any $x\in\mathcal{L}_a$, where again on the meta-level we have invoked the classical Bayes rule. Requiring that the "probability of probabilities" is not affected by the presence or absence of the filter,

$$\mathrm{prob}(\rho\,|\,\pi_b)=\mathrm{prob}(\rho)\ ,\tag{15}$$

and using $\mathrm{prob}(x\,|\,\rho,\pi_b)=\pi_b\rho\,(x)$ we find that

$$\mathrm{prob}(x\,|\,\pi_b)=\mathrm{prob}(x\mid\mathrm{prob}(\rho)\cdot\pi_b\rho+\mathrm{prob}(\sigma)\cdot\pi_b\sigma)\ ,\tag{16}$$

i.e., the map π_b is linear:

$$\pi_b(u\rho+v\sigma)=u\,\pi_b\rho+v\,\pi_b\sigma\ .\tag{17}$$

C. Transformations

Besides filtering, a second important class of experiments are "transformations" that do not involve the testing of any hypothesis, and whose effect amounts to a mere consistent relabeling of hypotheses; here "consistency" means that logical implications must be preserved. Such consistency-preserving transformations form a group $\mathcal{G}_a$ of automorphisms of $\mathcal{P}_a$ and $\mathcal{L}_a$, respectively, that satisfy

$$(g(\rho))(x) = \rho(g^{-1}(x)) \tag{18}$$

and

$$x \subseteq y \;\Leftrightarrow\; g(x) \subseteq g(y) \;. \tag{19}$$

The two types of experiment, (irreversible) filtering and (reversible) transformation, may be combined. Their order of execution can be exchanged provided the hypothesis being tested is subjected to relabeling, too:

$$g \circ \pi_b = \pi_{g(b)} \circ g \;. \tag{20}$$

If a system is described by a mixture of two probability distributions ρ, σ then, by a now familiar line of reasoning, transformation with $g \in \mathcal{G}_a$ leads to a posterior probability

$$\begin{aligned} \mathrm{prob}(x\,|\,g) \;&=\; \mathrm{prob}(x\,|\,\rho, g)\cdot\mathrm{prob}(\rho\,|\,g) + \\ &\quad +\mathrm{prob}(x\,|\,\sigma, g)\cdot\mathrm{prob}(\sigma\,|\,g) \end{aligned} \tag{21}$$

for any $x \in \mathcal{L}_a$. Requiring that the "probability of probabilities" is not affected by group action,

$$\mathrm{prob}(\rho\,|\,g) = \mathrm{prob}(\rho) \;, \tag{22}$$

and using $\mathrm{prob}(x\,|\,\rho, g) = (g(\rho))\,(x)$ we find that

$$\mathrm{prob}(x\,|\,g) = \mathrm{prob}(x \mid \mathrm{prob}(\rho)\cdot g(\rho) + \mathrm{prob}(\sigma)\cdot g(\sigma)) \;, \tag{23}$$

hence transformations are linear on $\mathcal{P}_a$:

$$g(u\rho + v\sigma) = u\, g(\rho) + v\, g(\sigma) \;. \tag{24}$$

III. EVIDENCE AS THE SOLE PARAMETER

Having ascertained the truth of a certain hypothesis a, the maximum amount of *additional* evidence that can still be garnered from a most refined experiment equals the maximum number of alternative refinements of a; it shall be denoted by

$$d(a) := \max \#\{b_i \mid \{b_i\} \prec a,\; b_i \neq \emptyset\} \tag{25}$$

and has the obvious properties

$$x \subseteq y \;\Rightarrow\; d(x) \leq d(y) \;, \tag{26}$$

$$d(x) = 0 \;\Leftrightarrow\; x = \emptyset \;. \tag{27}$$

Furthermore, it is group-invariant,

$$d(g(x)) = d(x) \;\forall\; x \in \mathcal{L}_a,\; g \in \mathcal{G}_a \;. \tag{28}$$

We are concerned with situations in which this maximum evidence is finite.

The above definition can be extended to probability distributions. For any probability distribution $\rho \in \mathcal{P}_a$ we first define its "support" $\mathrm{supp}(\rho)$ as the unique hypothesis in $\mathcal{L}_a$ for which

$$\mathrm{supp}(\rho) \subseteq x \;\Leftrightarrow\; \pi_x\rho = \rho \;. \tag{29}$$

The support transforms in a covariant fashion,

$$\mathrm{supp}(g(\rho)) = g(\mathrm{supp}(\rho)) \;, \tag{30}$$

and after filtering is constrained to be a refinement of the hypothesis just verified,

$$\mathrm{supp}(\pi_x\rho) \subseteq x \;, \tag{31}$$

with strict inequality if and only if x has some non-absurd refinement whose probability vanishes:

$$\mathrm{supp}(\pi_x\rho) \subset x \;\Leftrightarrow\; \exists\, y \subseteq x\,,\; y \neq \emptyset :\; \rho(y) = 0 \;. \tag{32}$$

We shall then define

$$d(\rho) := d(\mathrm{supp}(\rho)) \;. \tag{33}$$

Like its counterpart for hypotheses it is group-invariant, and it has the analogous properties

$$\rho \leq \sigma \;\Rightarrow\; d(\rho) \leq d(\sigma) \;, \tag{34}$$

$$d(\rho) = 0 \;\Leftrightarrow\; \rho = 0 \;. \tag{35}$$

Filtering generally produces new evidence and hence leads to a narrowing of probability distributions,

$$d(\pi_x\rho) \leq d(\rho) \;, \tag{36}$$

even though it is *not* necessarily $\mathrm{supp}(\pi_x\rho) \subseteq \mathrm{supp}(\rho)$.

We require that the "maximum available evidence" be the only parameter of the theory. This requirement has several important ramifications. To begin with, a hypothesis a can be decomposed into ever more accurate alternative refinements in an iterative, tree-like fashion by first identifying some initial complete set of alternative refinements, then decomposing each of these refinements into a further complete set of alternative refinements, and so on until this process comes to a halt because hypotheses cannot be refined any further. The absence of other parameters implies that regardless of the precise path chosen to arrive at such a maximal decomposition, the total number of outermost branches at the end of the process must always be the same and equal to the maximum evidence $d(a)$; which entails

$$\{b_i\} \prec a \;\Rightarrow\; d(a) = \sum_i d(b_i) \;. \tag{37}$$

Furthermore, whenever two hypotheses are at the same level of coarse-graining, $d(a) = d(b)$, then the corresponding substructures must be isomorphic: $\mathcal{L}_a \sim \mathcal{L}_b$ and $\mathcal{P}_a \sim \mathcal{P}_b$. The latter therefore form an equivalence class that depends on the maximum evidence only, and that we shall denote by $\mathcal{L}(d)$ and $\mathcal{P}(d)$, respectively. Likewise the associated structure group, too, depends on the maximum evidence only and shall be denoted by $\mathcal{G}(d)$.

Finally, any hypothesis x can have only one group-invariant property: its level of coarse-graining, $d(x)$. As long as they are at the same level of coarse-graining, two hypotheses can always be transformed into one another by some consistent relabeling,

$$d(x) = d(y) \;\Rightarrow\; \exists\, g \in \mathcal{G}(d) :\; y = g(x) \tag{38}$$

for any $x, y \in \mathcal{L}(d)$. Thus the set of all hypotheses at the same level of coarse-graining k,

$$\mathcal{M}_k(d) := \{x \in \mathcal{L}(d) \mid d(x) = k,\, k \leq d\} \;, \tag{39}$$

constitutes a homogeneous space on which $\mathcal{G}(d)$ acts transitively. The stability group of any $y \in \mathcal{M}_k(d)$ equals the product of $\mathcal{G}(k)$ acting on its substructure $\mathcal{L}_y \sim \mathcal{L}(k)$, and $\mathcal{G}(d-k)$ acting on

$$\{x \in \mathcal{L}(d) \mid x \perp y\} \sim \mathcal{L}(d-k) \;; \tag{40}$$

hence the set $\mathcal{M}_k(d)$ can be written as the quotient

$$\mathcal{M}_k(d) \sim \mathcal{G}(d)/\mathcal{G}(k) \otimes \mathcal{G}(d-k) \;. \tag{41}$$

This result can be generalised to complete sets of alternative refinements. The set

$$\mathcal{M}_{\{k_i\}}(d) := \{\, \{x_i\} \prec I_d \mid d(x_i) = k_i, \sum_i k_i = d\} \;, \tag{42}$$

where we have defined I_d as the maximal element of $\mathcal{L}(d)$ with $d(I_d) = d$, again constitutes a homogeneous space on which $\mathcal{G}(d)$ acts transitively. The stability group of any $\{y_i\} \in \mathcal{M}_{\{k_i\}}(d)$ now equals the product of all $\mathcal{G}(k_i)$ acting on the respective substructures $\mathcal{L}_{y_i} \sim \mathcal{L}(k_i)$; hence

$$\mathcal{M}_{\{k_i\}}(d) \sim \mathcal{G}(d)/\bigotimes_i \mathcal{G}(k_i) \;. \tag{43}$$

IV. DIMENSIONAL ANALYSIS

A. Preliminaries

The set of probability distributions $\mathcal{P}(d)$, the structure group $\mathcal{G}(d)$ and the set of hypotheses $\mathcal{M}_k(d)$ may be discrete or continuous. In case they are continuous we shall denote the dimensions of the respective manifolds by

$$P(d) := \dim \mathcal{P}(d) \;, \tag{44}$$

$$G(d) := \dim \mathcal{G}(d) \;, \tag{45}$$

$$M_k(d) := \dim \mathcal{M}_k(d) \;, \tag{46}$$

where the quotient representation (41) immediately implies the relation

$$M_k(d) = G(d) - G(k) - G(d-k) \;. \tag{47}$$

In the trivial case $d = 1$ there is only a single hypothesis, and any (non-normalised) probability distribution is uniquely specified by the probability of this single hypothesis being true. Therefore,

$$P(1) = 1 \;. \tag{48}$$

Classically, $\mathcal{M}_k(d)$ is a discrete set, $\mathcal{G}(d)$ is an equally discrete permutation group, and any (non-normalised) probability distribution is determined by d continuous parameters; hence

$$P_{\mathrm{cl}}(d) = d \;,\; G_{\mathrm{cl}}(d) = 0 \;,\; M_{k\,\mathrm{cl}}(d) = 0 \;. \tag{49}$$

In contrast, we are concerned here with situations in which hypotheses form a continuum. In the following we shall argue that then the only other consistent solution is

$$P(d) = d^2 \;,\; G(d) = d^2 \;,\; M_k(d) = 2k(d-k) \tag{50}$$

corresponding to $\mathcal{G}(d) \sim U(d)$. This will involve closer inspection of, and some additional assumptions pertaining to, (i) the preparation and (ii) composition of systems, as well as (iii) the continuity of probabilities.

B. Preparation

Any knowledge about a physical system, embodied in a probability distribution $\rho \in \mathcal{P}(d)$, is the result of a series of experiments or "preparation procedures" [21] applied to an initial state of total ignorance

$$\rho^{(0)}(x) := d(x)/d \;\forall\; x \in \mathcal{L}(d) \;. \tag{51}$$

This initial state of total ignorance is characterised by invariance under the structure group,

$$g(\rho^{(0)}) = \rho^{(0)} \;\forall\; g \in \mathcal{G}(d) \;, \tag{52}$$

in accordance with the "principle of indifference" [1].

Preparation procedures may be arbitrary combinations of (i) testing sets of mutually exclusive hypotheses; (ii) keeping or discarding the system, with respective probabilities that may depend on the outcome of the test; and (iii) transformations. In mathematical terms, for any $\rho \in \mathcal{P}(d)$ there exist sets of alternative refinements $\{b_i^{(\alpha)}\}$, sets of associated rescaling factors $\{\lambda_i^{(\alpha)}\}$ that reflect the respective probabilities of keeping or discarding the system, as well as transformations $\{g^{(\alpha)}\}$ such that

$$\begin{aligned} \rho \;=\; \ldots \circ g^{(\alpha)} \circ \left(\sum_i \lambda_i^{(\alpha)} \pi_{b_i^{(\alpha)}} \right) \circ \ldots \\ \ldots \circ g^{(1)} \circ \left(\sum_j \lambda_j^{(1)} \pi_{b_j^{(1)}} \right) \rho^{(0)} \;. \end{aligned} \tag{53}$$

Using linearity and the exchange rule (20) all transformations can be shifted to the right and absorbed in $\rho^{(0)}$, leaving behind only filters (pertaining to transformed sets $\{\tilde{b}\}$ of alternative refinements) and rescaling factors. In particular, one can define a sequence of posterior probability distributions

$$\rho^{(\alpha)} = \left(\sum_i \lambda_i^{(\alpha)} \pi_{\tilde{b}_i^{(\alpha)}} \right) \rho^{(\alpha-1)} \quad , \quad \alpha \geq 1 \tag{54}$$

after the α-th preparation procedure, that eventually terminates to yield ρ.

The left-hand side of the above iteration is some point on the manifold $\mathcal{P}(d)$, hence specified by $P(d)$ real parameters. The right-hand side, on the other hand, is uniquely specified by defining (i) the set $\{\tilde{b}_i\}$ of alternative refinements and (ii) for each refinement $\tilde{b}_j$, if ascertained, an associated posterior distribution in $\mathcal{P}_{\tilde{b}_j}$. Let $k_i := d(\tilde{b}_i)$ and hence $\{\tilde{b}_i\} \in \mathcal{M}_{\{k_i\}}(d)$. Then due to the quotient representation (43) the set of alternative refinements is specified by

$$\dim \mathcal{M}_{\{k_i\}}(d) = G(d) - \sum_i G(k_i) \tag{55}$$

real parameters; and a posterior distribution in $\mathcal{P}_{\tilde{b}_j} \sim \mathcal{P}(k_j)$ is specified by $P(k_j)$ real parameters. Equating the total number of parameters on the left-hand side and on the right-hand side of the iteration equation then yields a first constraint on the dimensions:

$$P(d) = G(d) - \sum_i G(k_i) + \sum_i P(k_i) \ , \ \sum_i k_i = d \ . \tag{56}$$

In the special case $k_i = 1$ for all i one obtains, using $P(1) = 1$,

$$G(d) = P(d) + (G(1) - 1) \cdot d \ . \tag{57}$$

C. Composition

Let a system be composed of two subsystems with respective maximum evidence $d^{(1)}$, $d^{(2)}$ and complete sets of alternative refinements $\{x_i^{(1)}\} \prec I_{d^{(1)}}$, $\{x_j^{(2)}\} \prec I_{d^{(2)}}$. Then the combined hypotheses $\{(x_i^{(1)}, x_j^{(2)})\}$—meaning "hypothesis $x_i^{(1)}$ pertaining to system 1 *and* hypothesis $x_j^{(2)}$ pertaining to system 2"—constitute a complete set of alternative refinements in the composite system. (Here the Boolean operation "and" is used in a perfectly classical sense since the two hypotheses refer to different subsystems and are thus jointly decidable; whereas for more general settings we carefully refrain from defining any of the conventional Boolean operations.) If the hypotheses about the subsystems are "most refined" then so are the combined hypotheses about the composite system,

$$d(x_i^{(1)}) = d(x_j^{(2)}) = 1 \ \Rightarrow \ d((x_i^{(1)}, x_j^{(2)})) = 1 \ ; \tag{58}$$

which implies that the maximum evidence about the composite system is the product $d^{(1)} \cdot d^{(2)}$.

Probability distributions for the two subsystems are specified by $P(d^{(1)})$ or $P(d^{(2)})$ real parameters, respectively. This means that there is a set of $P(d^{(1)})$ (not necessarily mutually exclusive) hypotheses $\{b_i^{(1)}\}$, and likewise a set of $P(d^{(2)})$ hypotheses $\{b_j^{(2)}\}$, such that the probabilities for these selected hypotheses uniquely determine the full distribution in the respective subsystem. Then for the composite system the full probability distribution is uniquely specified by the $P(d^{(1)}) \cdot P(d^{(2)})$ combined hypotheses $\{(b_i^{(1)}, b_j^{(2)})\}$; i.e., $P(d^{(1)}d^{(2)}) = P(d^{(1)})P(d^{(2)})$. Given $P(1) = 1$ this yields a second constraint on the dimensions [20]:

$$P(d) = d^{\mu} \quad , \quad \mu \in \mathbf{N} \ . \tag{59}$$

A similar line of reasoning can be applied to the composition of transformations. Isolated transformations of the two subsystems are specified by $G(d^{(1)})$ or $G(d^{(2)})$ real parameters, respectively. Hence, assuming the structure groups to be Lie groups, there are associated Lie algebras with $G(d^{(1)})$ generators $\{X_i^{(1)}\}$ and $G(d^{(2)})$ generators $\{X_j^{(2)}\}$, respectively. Then for the composite system there must be a larger Lie algebra whose generators are isomorphic to the $G(d^{(1)}) \cdot G(d^{(2)})$ pairs $\{(X_i^{(1)}, X_j^{(2)})\}$. This implies $G(d^{(1)}d^{(2)}) = G(d^{(1)})G(d^{(2)})$ and thus a third constraint on the dimensions:

$$G(d) = 0 \quad \text{or} \quad G(d) = d^{\nu} \quad , \quad \nu \in \mathbf{N} \ . \tag{60}$$

D. Continuity

We require that probabilities change under transformations in a continuous fashion, where "continuity" shall be defined as follows. Assuming that the structure group $\mathcal{G}(d)$ is a Lie group and hence endowed with a group-invariant distance measure then it is possible to define, for any (infinitesimal) $\delta > 0$, an (infinitesimal) neighborhood of the identity element $1_{\mathcal{G}}$

$$\mathcal{G}_\delta(d) := \{g \in \mathcal{G}(d) \mid \mathrm{dist}(g, 1_{\mathcal{G}}) < \delta\} \ . \tag{61}$$

Given a probability distribution $\rho \in \mathcal{P}(d)$, all refinements of its support have non-vanishing probabilities that are greater than or equal to

$$\epsilon(\rho) := \min\{\rho(x) \mid x \subseteq \mathrm{supp}(\rho) \ , \ x \neq \emptyset\} > 0 \ . \tag{62}$$

Now "continuity" means that probabilities that were initially greater than zero not suddenly jump to zero upon an infinitesimal transformation; in more rigorous mathematical terms,

$$\forall\, \epsilon(\rho) > 0 \ \exists\, \delta > 0 : \ g(\rho)\,(x) > 0 \ \ \forall \ \ x \subseteq \mathrm{supp}(\rho) \ , \ x \neq \emptyset \ , \quad g \in \mathcal{G}_\delta(d) \ . \tag{63}$$

By virtue of Eq. (32) this is equivalent to requiring

$$\mathrm{supp}\left[\pi_{\mathrm{supp}(\rho)} g(\rho)\right] = \mathrm{supp}(\rho) \;\forall\; g \in \mathcal{G}_\delta(d) \;. \tag{64}$$

In the remainder of this section we shall always assume that we are in the infinitesimal neighborhood $g \in \mathcal{G}_\delta(d)$.

For further analysis we introduce an arbitrary auxiliary hypothesis b,

$$\mathrm{supp}(\rho) \subseteq b \subseteq I_d \;, \tag{65}$$

where the respective levels of coarse-graining $k := d(\mathrm{supp}(\rho))$, $l := d(b)$ and $d \equiv d(I_d)$ satisfy

$$k \leq l \leq d \;. \tag{66}$$

Moreover, we define three additional auxiliary hypotheses z, $b\backslash z$ and b_z^* as follows:

$$z := \mathrm{supp}\left[\pi_b g(\rho)\right] \subseteq b \;, \tag{67}$$

$$\{b\backslash z, z\} :\prec b \tag{68}$$

and

$$\{b_z^*, b\backslash z\} :\prec I_d \;. \tag{69}$$

These definitions imply

$$z,\; g(\mathrm{supp}(\rho)) \subseteq b_z^* \;. \tag{70}$$

Within the continuous region the associated levels of coarse-graining take the values

$$d(z) = k \;,\; d(b\backslash z) = l - k \;,\; d(b_z^*) = d - l + k \;. \tag{71}$$

The proofs of (70) and (71) are given in the appendix.

As $\mathrm{supp}(\rho)$ and z are both refinements of b and have the same level of coarse-graining k, they are both elements of the set

$$\{x \in \mathcal{L}_b \mid d(x) = k,\; k \leq l\} \sim \mathcal{M}_k(l) \;; \tag{72}$$

hence given b, the hypothesis z is uniquely specified by $M_k(l)$ real parameters. Likewise, z and $g(\mathrm{supp}(\rho))$ are both refinements of b_z^*, again at the same level of coarse-graining k, and thus elements of the set

$$\{x \in \mathcal{L}_{b_z^*} \mid d(x) = k,\; k \leq (d - l + k)\} \sim \mathcal{M}_k(d - l + k) \;; \tag{73}$$

so given both b and z, and hence b_z^*, the transformed support $g(\mathrm{supp}(\rho))$ is uniquely specified by $M_k(d-l+k)$ real parameters. Therefore the total number of parameters needed to specify $g(\mathrm{supp}(\rho))$ is the sum $M_k(l) + M_k(d - l + k)$, which must equal the number of parameters that would have been needed *without* the above auxiliary construction:

$$M_k(d) = M_k(l) + M_k(d - l + k) \;. \tag{74}$$

In combination with Eq. (47) this implies the fourth and final constraint on the dimensions:

$$G(d) = \frac{G(2) - 2G(1)}{2} d(d-1) + G(1)\, d \;. \tag{75}$$

E. Summary

The four constraints (57), (59), (60) and (75) together with Eq. (47) permit only three solutions: (i) the "classical case" in which hypotheses constitute a discrete set, the structure group is equally discrete, and any (non-normalised) probability distribution is determined by d continuous parameters:

$$P_{\mathrm{cl}}(d) = d \;,\; G_{\mathrm{cl}}(d) = 0 \;,\; M_{k\,\mathrm{cl}}(d) = 0 \;; \tag{76}$$

(ii) a case in which the set of hypotheses is still discrete and probability distributions are still determined by d continuous parameters, yet there is a continuous group introducing non-trivial phases:

$$P_{\mathrm{sc}}(d) = d \;,\; G_{\mathrm{sc}}(d) = d \;,\; M_{k\,\mathrm{sc}}(d) = 0 \;, \tag{77}$$

corresponding to $\mathcal{G}(d) \sim U(1)^{\otimes d}$; we may think of this as a "semiclassical case"; and (iii) the only allowed case in which hypotheses form a continuum:

$$P_{\mathrm{qu}}(d) = d^2 \;,\; G_{\mathrm{qu}}(d) = d^2 \;,\; M_{k\,\mathrm{qu}}(d) = 2k(d-k) \;. \tag{78}$$

Given that $\mathcal{G}(d)$ must be a compact Lie group this leads to $\mathcal{G}(d) \sim U(d)$ [22]. This last case proves our original conjecture: Whenever hypotheses form a continuum but evidence is restricted to be finite, the *only* consistent framework for plausible reasoning is the complex Hilbert space framework of quantum theory.

One may wonder what happens if any of the constraints are relaxed. Table I gives an overview of our requirements for consistent reasoning, the associated dimensional constraints, and the additional cases allowed if a constraint is relaxed in isolation. The requirements pertaining to preparation and to the composition of states are not instrumental in—but perfectly consistent with—deriving the dimensionality of the structure group; without the preparation requirement, however, the connection is lost between the group dimension and the dimension of the state manifold. If, instead, the requirement pertaining to the composition of transformations is relaxed then on purely dimensional grounds one additional structure group becomes possible: $SO(d) \otimes SO(d)$. This new structure group leaves the dimensions of the various manifolds of hypotheses $\mathcal{M}_k(d)$ unchanged but changes their topology, e.g., $\mathcal{M}_1(2)$ becomes isomorphic to the surface of a torus rather than the surface of a sphere. The physical significance of such topologies that are not simply connected remains elusive. However, they might be in conflict with the requirement that the set of probability distributions be convex [20]. Finally, if the continuity requirement is relaxed in isolation then the dimensions of group and state manifold, while constrained to be equal, may be higher powers of d. Again it is not clear what the physical significance of such a behavior would be.

TABLE I: Overview of requirements for consistent reasoning, associated dimensional constraints, and additional cases allowed if a constraint is relaxed in isolation.

		Extra cases allowed if relaxed	
Requirement	Implied dimensional constraint	Dimensions	Structure group
Preparation	$G(d) = P(d) + (G(1) - 1)\, d$	$P(d) = d^\mu\ ,\ \mu \neq \nu$	—
Composition (states)	$P(d) = d^\mu$	—	—
Composition (transformations)	$G(d) = 0, d^\nu$	$G(d) = d(d-1)$	$SO(d) \otimes SO(d)$
Continuity	$G(d) = \frac{G(2)-2G(1)}{2} d(d-1) + G(1)\, d$	$G(d) = P(d) = d^\mu\ ,\ \mu \geq 3$	many

V. CONCLUSIONS

We have considered the non-classical situation in which hypotheses form a continuum, whereas the maximum available evidence is bounded from above by some finite integer d. Employing the basic notions of hypotheses, probabilities, filters and transformations, and invoking a small number of consistency requirements pertaining to the preparation and composition of systems, as well as to the continuity of probabilities, we have shown that then the group of consistency-preserving transformations must be isomorphic to $U(d)$. Our proof highlights the pivotal role played by the finite maximum evidence alias Hilbert space dimension d as the sole parameter of the theory, confirming an earlier intuition by Fuchs [23].

We have thus singled out complex Hilbert space as the *only* consistent framework for plausible reasoning. Quantum theory is indeed an "island in theoryspace" [24] distinguished by a high degree of internal consistency. In particular, alternative models in real [25] or quaternionic [26] Hilbert spaces that are allowed by traditional quantum logic [7] (but that have already run into difficulties for other reasons such as the lack of a de Finetti representation [27]) now seem very difficult to justify. We also note that nowhere did we make reference to specific length or energy scales; hence even though quantum phenomena are most prevalent in the microscopic world, there is nothing in the above line of argument that restricts it to that domain.

Once identified with quantum theory in complex Hilbert space, the various notions of statistical inference employed in this paper can be easily translated into the familiar language of conventional quantum theory; these correspondences are summarised in Table II. As is well known, quantum theory entails a number of counterintuitive features. We recall a few, using the terminology of this article: (i) The classical Boolean operations "and", "or" are not well defined for arbitrary hypotheses. Indeed, even though in certain special cases they are implicit in our above definitions of π_b, $\perp$ or $\prec$, we have avoided employing these notions in our line of argument. (ii) Some pairs of hypotheses are not jointly decidable, making quantum theory inherently probabilistic. (Two hypotheses $x, y \in \mathcal{L}(d)$ are said to be jointly decidable if there is a complete set of alternative refinements $\{b_i\}_{i\in I} \prec I_d$ with subsets of the index set $I_x, I_y \subseteq I$ such that $\{b_i\}_{i\in I_x} \prec x$ and $\{b_i\}_{i\in I_y} \prec y$.) (iii) It is not possible to assign to all hypotheses a preexisting truth value, i.e., to mimic quantum theory with a hidden-variables theory [28].

Niels Bohr once remarked that physics in general, and quantum theory in particular, was to be regarded "not so much as the study of something a priori given" but rather as the development of "methods for ordering and surveying human experience" [29]. I hope this paper will have further corroborated the deep truth of this statement, provided we interpret "ordering and surveying human experience" as meaning "consistent reasoning about hypotheses pertaining to the physical world".

Acknowledgments

I thank Berndt Müller for critical reading of the manuscript and valuable feedback.

APPENDIX

1. Proof of Eq. (70)

That $z \subseteq b_z^*$ follows directly from their respective definitions. The second logical implication in Eq. (70) can be shown as follows. It is

$$\rho\left(g^{-1}(b) \backslash \mathrm{supp}\left(\pi_{g^{-1}(b)}\rho\right)\right) = 0 \tag{A.1}$$

and hence

$$\rho\left(g^{-1}(b) \backslash g^{-1}(z)\right) = 0\ , \tag{A.2}$$

which implies

$$g(\rho)\ (b \backslash z) = 0 \tag{A.3}$$

and further

$$g(\mathrm{supp}(\rho)) \perp b \backslash z\ . \tag{A.4}$$

This yields

$$g(\mathrm{supp}(\rho)) \subseteq b_z^*\ , \tag{A.5}$$

Q.E.D.

TABLE II: Correspondence between the terminology of statistical inference employed in this paper and the terminology of conventional quantum theory.

Statistical inference		Quantum theory	
Name	Symbol or relation	Name	Symbol or relation
Hypothesis	x	Projector	$\hat{P}_x$, x is subspace of Hilbert space
Probability distribution	ρ	Density matrix	$\hat{\rho}$
Probability	$\rho(x)$	Probability	$\mathrm{tr}(\hat{\rho}\hat{P}_x)$
Logical implication	$x \subseteq y$	—	$\hat{P}_x\hat{P}_y = \hat{P}_y\hat{P}_x = \hat{P}_x$
Filter	$\pi_b\rho$	—	$\hat{P}_b\hat{\rho}\hat{P}_b$
Contradiction	$x \perp y$	Orthogonality	$\hat{P}_x\hat{P}_y = \hat{P}_y\hat{P}_x = 0$
Complete set of alternative refinements	$\{b_i\} \prec a$	Orthogonal decomposition	$\hat{P}_a = \sum_i \hat{P}_{b_i}$
Transformation	$g(\rho)$	Unitary transformation	$\hat{U}\hat{\rho}\hat{U}^\dagger$
Level of coarse graining	$d(x)$	Dimension of subspace	$\mathrm{tr}(\hat{P}_x)$
Most refined hypothesis	$d(x) = 1$	1-dim. subspace (ray)	$\hat{P}_x = \|\chi\rangle\langle\chi\|$

2. Proof of Eq. (71)

Eq. (36) and group invariance imply

$$d(\pi_b g(\rho)) \leq d(g(\rho)) = d(\rho) \; ; \tag{A.6}$$

while $b \supseteq \mathrm{supp}(\rho)$ and the continuity condition (64) yield

$$d(\pi_b g(\rho)) \geq d(\pi_{\mathrm{supp}(\rho)} g(\rho)) = d(\rho) \; . \tag{A.7}$$

Together these inequalities give

$$d(\pi_b g(\rho)) = d(\rho) \tag{A.8}$$

and hence $d(z) = k$, Q.E.D.

[1] E. T. Jaynes, *Probability theory: the logic of science* (Cambridge University Press, 2003).
[2] D. S. Sivia, *Data analysis: a Bayesian tutorial* (Oxford University Press, 1996).
[3] R. T. Cox, Am. J. Phys. **14**, 1 (1946).
[4] C. M. Caves, C. A. Fuchs, and R. Schack, Phys. Rev. A **65**, 022305 (2002).
[5] M. Srednicki, Phys. Rev. A **71**, 052107 (2005).
[6] G. Birkhoff and J. v. Neumann, Ann. Math. **37**, 823 (1936).
[7] J. M. Jauch, *Foundations of quantum mechanics* (Addison-Wesley, 1968).
[8] V. S. Varadarajan, *Geometry of quantum theory* (Springer, 1985), 2nd ed.
[9] D. W. Cohen, *An introduction to Hilbert space and quantum logic* (Springer, 1989).
[10] C. Piron, Helv. Phys. Acta **37**, 439 (1964).
[11] R. Schack, T. A. Brun, and C. M. Caves, Phys. Rev. A **64**, 014305 (2001).
[12] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information* (Cambridge University Press, 2000).
[13] A. Steane, Rept. Prog. Phys. **61**, 117 (1998).
[14] A. Peres and D. R. Terno, Rev. Mod. Phys. **76**, 93 (2004).
[15] M. Keyl, Phys. Rep. **369**, 431 (2002).
[16] R. Clifton, J. Bub, and H. Halvorson, Found. Phys. **33**, 1561 (2003).
[17] J. B. Hartle, Am. J. Phys. **36**, 704 (1968).
[18] A. Einstein, B. Podolsky, and N. Rosen, Phys. Rev. **47**, 777 (1935).
[19] N. Bohr, Phys. Rev. **48**, 696 (1935).
[20] L. Hardy, quant-ph/0101012.
[21] A. Peres, *Quantum theory: concepts and methods* (Kluwer Academic Publishers, 1995).
[22] A. O. Barut and R. Raczka, *Theory of group representations and applications* (World Scientific, 1986), 2nd ed.
[23] C. A. Fuchs, quant-ph/0205039.
[24] S. Aaronson, quant-ph/0401062.
[25] E. C. G. Stueckelberg, Helv. Phys. Acta **33**, 727 (1960).
[26] D. Finkelstein, J. M. Jauch, S. Schiminovich, and D. Speiser, J. Math. Phys. **3**, 207 (1962).
[27] C. M. Caves, C. A. Fuchs, and R. Schack, J. Math. Phys. **43**, 4537 (2002).
[28] N. D. Mermin, Rev. Mod. Phys. **65**, 803 (1993).
[29] N. Bohr, *Essays 1958-1962 on atomic physics and human knowledge* (Wiley, 1963).