frontiers | Frontiers in Cognition

TYPE Hypothesis and Theory
PUBLISHED 06 August 2025
DOI 10.3389/fcogn.2025.1623227

OPEN ACCESS

EDITED BY
Andrew Tolmie,
University College London, United Kingdom

REVIEWED BY
Emmanuel E. Haven,
Memorial University of Newfoundland, Canada
Pier Luigi Gentili,
Università degli Studi di Perugia, Italy

*CORRESPONDENCE
Bart Jacobs
✉ bart@cs.ru.nl

RECEIVED 05 May 2025
ACCEPTED 25 June 2025
PUBLISHED 06 August 2025

CITATION
Jacobs B (2025) Commutativity of probabilistic belief revision.
Front. Cognit. 4:1623227.
doi: 10.3389/fcogn.2025.1623227

Commutativity of probabilistic belief revision

Bart Jacobs*

iHub, Radboud University, Nijmegen, Netherlands

Bayesian updating, also known as belief revision or conditioning, is a core mechanism of probability theory, and of AI. The human mind is very sensitive to the order in which it is being "primed", but Bayesian updating works commutatively: the order of the evidence does not matter. Thus, there is a mismatch. This paper develops Bayesian updating as an explicit operation on (discrete) probability distributions, so that the commutativity of Bayesian updating can be clearly formulated and made explicit in several examples. The commutativity mismatch is underexplored, but plays a fundamental role, for instance in the move to quantum cognition.

KEYWORDS

Bayesian updating, multiset, commutativity, notation, cognition

1 Introduction

In mathematics an operation is called commutative if swapping its arguments does not change the outcome, as in addition of numbers: $n + m = m + n$. Also, actions can be called commutative when the effect does not depend on the order in which they are taken: if I first give someone n Euros and then another m Euros, the financial effect is the same as first giving you m and then n Euros. However, the emotional effect on the receiver's side may be quite different, for instance when n is much greater than m: first giving the higher amount n then m may lead to disappointment, whereas after first giving the lower amount m and then n the receiver may end up in a more positive mood.

Similar differences are well-known in human cognition, especially when one looks at how the mind processes (new) information—in what is called priming. The order of such priming (or updating) is highly relevant. Here is a simple example, involving two sentences p and q, about Alice and Bob, which will be presented below in two orders: (a) as p and q, and (b) as q and p. This makes no difference in (Boolean) logic, but watch carefully what effect the different orders has on your understanding of the situation.

(a) "Alice is sick" and "Bob visits Alice";
(b) "Bob visits Alice" and "Alice is sick."

In the first case (a) you may think that Bob is a nice guy, but maybe less so in the second case (b). It is surprising how quickly the human mind makes a (causal) connection. The strength of this effect depends on many factors, including one's own background (priors). Check for instance what the effect is of replacing in the above two sentences "sick" with "pregnant." This dependence on the order is called the order effect in Uzan (2023). It is the main motivation to switch to quantum logic in cognition theory (see e.g., Busemeyer and Bruza, 2012; Yearsley and Busemeyer, 2016; Gentili, 2021), since conjunction ("and") in quantum logic is not commutative.

This paper offers reflections on commutativity in probability theory, and in particular in probabilistic updating. Bayesian updating is introduced as an explicit operation that takes a probability distribution ω with some form of evidence p and produces a new,

updated distribution $\omega|_p$. This formulation generalizes the common approach. The (mathematical) details will appear later, but at this stage it is relevant to emphasize that this new formulation of Bayesian updating makes it possible to clearly express commutativity: for two pieces of evidence p and q one has:

$$\omega|_p|_q = \omega|_q|_p. \tag{1}$$

In words: updating a distribution ω first with p and then with q gives the same result as updating ω first with q and then with p. This commutativity cannot be expressed using the traditional notation $P(E \mid D)$ of conditional probabilities, since it leaves the distribution implicit.

To add some terminology: this updating is also called conditioning or belief revision. The distribution ω, before the update, is called the prior, and the distribution $\omega|_p$, after the update, is called the posterior. The task of computing the posterior distribution is called inference, or also (probabilistic) learning.

Probabilistic updating is an essential part of the ongoing AI-revolution, in various forms, via learning and training. Generative and conversational AI are becoming part of professional and private environments. If these tools are meant to behave like humans, their updating should be non-commutative, as illustrated above, with the different effects resulting from different orders (a) and (b). Bayesian updating, however, is commutative (Equation 1). Hence there is a mismatch [as also emphasized in Uzan (2023)]. One aim of this paper is increase the awareness of this gap, by making the commutativity of Bayesian updating explicit, in a new form (Equation 1).

The paper starts with some simple observations about lists and multisets. One can have a list of letters, say (a, b, c, a, c, b, a). In a list, elements can occur multiple times and the order of their occurrence matters. Notice that in a subset like $\{a, c\}$, elements can occur at most once, and their order does not matter. Mulitsets are "inbetween" lists and subsets: elements may occur multiple times, but their order does not matter. The latter property makes them relevant in the current context. Multisets are a highly undervalued datatype. The fact that they are so little used may be part of our poor understanding of the role of commutativity. We can already make a connection with what we saw above: updating a distribution with two pieces of evidence—as in Equation 1—should not be done with a list (p, q) or (q, p), but with a multiset of evidence containing both p and q, where the order does not matter. Section 2 below starts with an informal introduction to multisets and ends with some notation and definitions that are relevant in this setting.

Multisets form a natural preparation for (discrete probability) distributions, in Section 3. Distributions keep count of elements via probabilities or weights (in the unit interval $[0, 1]$) that add up to one. Multisets can be turned into "fractional" distributions via normalization. A basic fact is that every distribution can be obtained as limit of such fractional distributions, just like every real number can be obtained as a limit of fractions. Mathematically, this is described as: the set $\mathcal{D}(X)$ of distributions on a finite set X is a compact complete metric space, with (normalized) multisets as dense subset, see Theorem 1. This denseness formalizes the "frequentist" perspective on probability distributions, as results of long-term experiments.

Section 4 introduces Bayesian updating $\omega|_p$ in concrete form, shows that traditional notation $P(E \mid D)$ is a special case, and illustrates usage of updates $\omega|_p$ in two examples. The first one is a rather straightforward application, where a prior bird distribution is updated after a bird count. If there is another bird count one year later, the bird distribution can be updated once again. It turns out that eventhough the counts are chronologically ordered, this order is irrelevant for the distribution updates. The second example is more challenging. It involves various inference questions about the sex and ages of children in a family (with a twin), after specific observations. The possibilities for the offspring are represented as multisets, based on Section 2. The prior distribution in this case is a distribution over these multisets. Again, the order of the multiple updates does not matter. This basic fact is proven in general form at the end of this section, see Proposition 1.

The commutativity of Bayesian updating may be seen as folklore knowledge but it is hardly made explicit in the literature. One reason is that the standard formulation of conditional probabilities $P(E \mid D)$ does not lend it self to a commutativity result, as above in Equation 1, since it hides the distribution, assuming there is only one implicit distribution. Hence one cannot express facts about different distributions via traditional notation. Our formulation of Bayesian updating $\omega|_p$ as an operation on distributions thus has advantages—as hopefully also becomes clear from the illustrations in Section 4. The Appendix derives the update formulation $\omega|_p$ from the traditional formulation via Kadison duality. This is a new result. The derivation is mathematically sophisticated and is not necessary for the main line of the paper. This line is part of a new approach to probability theory using the language and methods of category theory. In the body of the paper the role of category theory remains in the background and no prior knowledge of that field is required.

Thus, this paper's contributions lie in putting the spotlight on the commutativity of Bayesian updating, in a new form (Equation 1), via a new formulation $\omega|_p$ of this update mechanism, which is both given in concrete form and derived from a fundamental duality result. At the same time, the paper provides a gentle introduction to a new approach to the area, in which multisets and explicitly written distributions play a central role.

2 Multisets, with multiplicities of elements

Suppose you check how much money you have in your pocket and you find that you have three 2-Euro coins (2€) and two 1-Euro coins (1€). How would you describe this handful of coins mathematically? It is not a subset of coins $\{(2€), (1€)\}$, since such a subset ignores the multiplicities of the coins that you have. One can describe the contents of your pocket as a list, for instance as, $((2€), (1€), (2€), (1€), (2€))$, when you lay them out in your hand. But the order of this list is arbitrary and does not reflect your answer: I have three of (2€) and two of (1€).

The proper way to capture the situation mathematically is via a *multiset*. It can be understood as a subset in which elements may occur multiple times, or as a list in which the order does not matter. Unfortunately, there is no established notation for multisets. We

use kets $|-\rangle$, where we put the elements of the multiset inside the ket and their multiplicity in front. Thus, the coins in your pocket, are properly described as multiset:

$$3\left|\text{2€}\right\rangle + 2\left|\text{1€}\right\rangle.$$

These kets $|-\rangle$ are borrowed from quantum theory. They have no mathematical meaning here and are used only to separate the elements of a multiset from their multiplicities.

As a practical example, consider the outcome of an election, say between two candidates Alice (A) and Bob (B). The outcome may be written as a multiset $55|A\rangle + 45|B\rangle$, indicating that 55 votes are for Alice and 45 votes for Bob. Describing the outcome of the election as a list of length 100, say with the chronological order of casted votes, is a bad idea, for two reasons: the list does not immediately tell what the outcome is, and the list may leak information about who voted for whom—when the order of the voters is recorded.

In which ways can you break a note of 10 Euro into coins of 2 and 1 Euros? The six options can be described as multisets:

$$\begin{array}{ccc} 5\left|\text{2€}\right\rangle & 4\left|\text{2€}\right\rangle + 2\left|\text{1€}\right\rangle & 3\left|\text{2€}\right\rangle + 4\left|\text{1€}\right\rangle \\ 2\left|\text{2€}\right\rangle + 6\left|\text{1€}\right\rangle & 1\left|\text{2€}\right\rangle + 8\left|\text{1€}\right\rangle & 10\left|\text{1€}\right\rangle. \end{array}$$

When we are interested in the ways to break the note of 10, we do not care about the order of the coins. When we do describe the break-up options as lists we end up with 10 different lists of coins.

Multisets are a useful "datatype," in the language of computer science, for keeping counts of elements. However, multisets are often not recognized or expounded as such. For instance, in mathematics, the solutions of a polynomial form a multiset, and not a set, since solutions may occur multiple times. For example, the multiset of solutions of the polynomial $x^3 - 7x^2 + 16x - 12 = (x-2)(x-2)(x-3)$ takes the form $2|2\rangle + 1|3\rangle$, since the number 2 occurs twice as solution and 3 once. Similarly, the eigenvalues of a matrix form a multiset. In the notation $2|2\rangle + 1|3\rangle$ the kets play a useful role, since they make clear which numbers are in the multiset and which numbers are the corresponding multiplicities.

Consider the following basic question. A friend of mine has three children, but I don't know if they are girls (G) or boys (B). How many offspring options are there? Many people will quickly say: *eight*, namely:

$$\begin{array}{cccc} G,G,G & G,G,B & G,B,G & B,G,G \\ G,B,B & B,G,B & B,B,G & B,B,B. \end{array} \tag{2}$$

One can also say: there are *four* options, namely with three, two, one, or zero girls. These four options are described as multisets:

$$3|G\rangle \quad 2|G\rangle + 1|B\rangle \quad 1|G\rangle + 2|B\rangle \quad 3|B\rangle. \tag{3}$$

The eight list options in Equation 2 only make sense if there is an order—*e.g.* by ascending or descending age—but no order was specified in the question. Hence the multiset answer, with four options, seems most natural. It abstracts away from any ordering of the children.

In the next section we illustrate how multisets form the basis for discrete probability theory. Indeed, probabilities naturally arise from counting, possibly in a limit process. Urns filled with balls of different colors form a basic model in probability theory (see e.g. Johnson and Kotz, 1977; Mahmoud, 2008; Ross, 2018) and many other references. Here, such an urn is identified with a multiset over the set of colors (see Figure 1). The multiplicities of the different colors determine the probability of drawing a ball of a particular color. For instance, in this case, the probability of drawing a single red ball is $\frac{4}{9}$. A general draw from such an urn is also a multiset, as on the right in Figure 1.

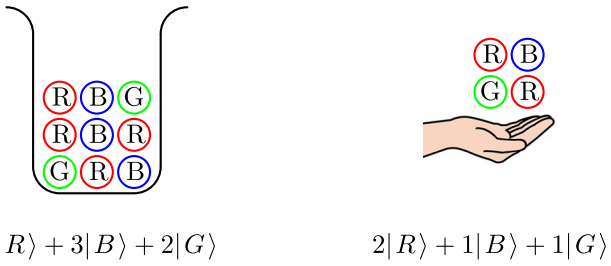

FIGURE 1
An urn filled with colored balls on the left, written as multiset, and a possible draw from this urn on the right, also written as a multiset. One can ask what is the probability of this draw from the urn. This depends on the mode of drawing: draw-and-delete, draw-and-replace, draw-and-duplicate (see Jacobs, 2022, 2025) for details.

These introductory observations illustrate that multisets are useful mathematical abstractions for keeping count of multiple occurrences of different objects (or data items). We have mentioned (Equation 1) that the order of data in Bayesian updating is irrelevant. This implies that data in Bayesian learning is best organized as a multiset. Indeed, a histogram of data—with heights in natural numbers—is another example of a multiset.

The next definition fixes the notation and terminology that we shall use in the sequel. In this setting, a multiset involves only finitely many elements from a given set. The multiplicities are natural numbers. One could also allow non-negative real numbers as multiplicities, but we don't need such generalizations. Here and in the sequel we shall use the sign $:=$ for definitions.

Definition 1. Let X be an arbitrary set.

1. A *multiset* over X is a finite formal sum of the form:

$$n_1|x_1\rangle + \cdots + n_k|x_k\rangle \text{ with } \begin{cases} \text{multiplicities } n_1, \ldots, n_k \in \mathbb{N} \\ \text{elements } x_1, \ldots, x_k \in X. \end{cases}$$

Alternatively, a multiset over X may be described as a function $\varphi\colon X \to \mathbb{N}$ with finite support $supp(\varphi) := \{x \in X \mid \varphi(x) \neq 0\}$.

2. The *size* $\|\varphi\| \in \mathbb{N}$ of a multiset φ is its total number of elements, including multiplicities. Explicitly, both in ket and function notation:

$$\Big\| \textstyle\sum_i n_i|x_i\rangle \Big\| := \sum_i n_i \text{ and } \big\| \varphi \big\| := \sum_{x \in supp(\varphi)} \varphi(x) = \sum_{x \in X} \varphi(x).$$

3. We shall write $\mathcal{M}(X)$ for the set of all multisets over X. This operation $\mathcal{M}$ is functorial: it works not only on sets X but also on functions $f\colon X \to Y$, and then yields a function $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ given by:

$$\mathcal{M}(f)\Big(\textstyle\sum_i n_i|x_i\rangle\Big) := \sum_i n_i|f(x_i)\rangle. \tag{4}$$

Notice that in this definition the set X may be infinite, but a multiset over X has only finitely many elements from X, in its support. We freely switch between the ket and function notation for multisets and use whichever is most convenient in a particular situation.

The ket notation involves a formal sum, with some conventions: (1) terms $0|x\rangle$ with multiplicity zero may be omitted; (2) a sum $n|x\rangle + m|x\rangle$ is the same as $(n+m)|x\rangle$; (3) the order and any round brackets in a sum do not matter. Thus, for instance, there is an equality of multisets:

$$2|a\rangle + \big(5|b\rangle + 0|c\rangle\big) + 4|b\rangle = 9|b\rangle + 2|a\rangle.$$

These conventions are especially relevant for functoriality in item 3, since multiple elements x, x' may be mapped to the same outcome $f(x) = f(x')$. We use this functoriality especially for projection functions $\pi_i\colon X_1 \times X_2 \to X_i$. It then yields marginalisaiton.

As briefly discussed in the introduction, it is illuminating to compare multisets to the datatypes of lists and subsets.

	lists	multisets	subsets
order of elements matters	+	-	-
multiplicity of elements matters	+	+	-

If we write $\mathcal{L}(X)$ and $\mathcal{P}(X)$ for the sets of finite lists and of subsets over a set X, then we can form a diagram in which multisets sit inbetween lists and subsets:

$$\mathcal{L}(X) \xrightarrow[\text{forget order}]{acc} \mathcal{M}(X) \xrightarrow[\text{forget multiplicity}]{supp} \mathcal{P}(X) \tag{5}$$

The function *acc* performs accumulation, via $acc(x_1, \ldots, x_n) := 1|x_1\rangle + \cdots + 1|x_n\rangle$. It counts the occurrences of elements in a list, for instance in:

$$acc(c, b, a, a, a, b, c) = 3|a\rangle + 2|b\rangle + 2|c\rangle.$$

Similarly, the accumulations of the lists of children in Equation 2 yields the multisets of children in Equation 3.

The support map *supp* in Equation 5 sends a multiset on a set to its subset of elements, see Definition 1 1. In the example, $supp(3|a\rangle + 2|b\rangle + 2|c\rangle) = \{a, b, c\}$. As an aside for the mathematically oriented reader, both *acc* and *supp* preserve the monoid structures on these data types and they both are maps of monads. They are fundamental, well-behaved mappings.

3 Distributions, with probabilities of elements

This section introduces finite discrete probability distributions, using ket notation as for multisets. It is shown that multisets give rise to "fractional" distributions, via normalization, and that each distribution is in fact a limit of such fractional distributions.

A coin has two sides, namely head (H) and tails (T). A fair coin assigns a probability of $\frac{1}{2}$ to both sides. In ket notation we write it as on the left below.

$$\tfrac{1}{2}|H\rangle + \tfrac{1}{2}|T\rangle \qquad\qquad \tfrac{51}{100}|H\rangle + \tfrac{49}{100}|T\rangle.$$

On the right, above, there is an almost fair coin, with a slight bias toward head. Characteristically for distributions, the numbers before the kets must be probabilties from the unit interval $[0, 1]$ that add up to one.

Consider the urn multiset $\upsilon = 4|R\rangle + 3|B\rangle + 2|G\rangle$ from Figure 1, with size $\|\upsilon\| = 9$. The probability of drawing a red ball is $\frac{4}{9} = \frac{\upsilon(R)}{\|\upsilon\|}$. These draw probabilities arise via normalization of the urn-as-multiset, for which we use a function *flrn*, as in:

$$\begin{aligned} flrn(\upsilon) &= \frac{\upsilon(R)}{\|\upsilon\|}|R\rangle + \frac{\upsilon(B)}{\|\upsilon\|}|B\rangle + \frac{\upsilon(G)}{\|\upsilon\|}|G\rangle \\ &= \frac{4}{9}|R\rangle + \frac{3}{9}|B\rangle + \frac{2}{9}|G\rangle. \end{aligned}$$

The latter distribution captures the probabilities of drawing a ball of a particular color from the urn υ. The function *flrn* will be defined in general form below. It turns a (non-empty) multiset into a distribution. It learns this distribution by counting, so *flrn* is used as abbreviation of frequentist learning.

As argued in Gigerenzer and Hoffrage (1995), people are in general not very good at probabilistic (esp. Bayesian) reasoning, but they fare better at reasoning with what are called "frequency formats" but which are in fact multisets. The information that there is a 0.04 probability of getting a disease can be captured in a distribution $\frac{1}{25}|D\rangle + \frac{24}{25}|D^\perp\rangle$, where $D^\perp$ stands for no-disease. This appears more difficult to grasp than the information that 4 out of 100 people get the disease, as captured by the multiset $4|D\rangle + 96|D^\perp\rangle$. Applying frequentist learning *flrn* to the latter multiset gives the disease distribution.

Definition 2. Let X, Y be arbitrary sets.

1. A *distribution* on X is a formal finite convex sum $r_1|x_1\rangle + \cdots + r_n|x_n\rangle$ with elements $x_i \in X$ and associated probabilities $r_i \in [0, 1]$ satisfying $\sum_i r_i = 1$.

   Alternatively, a distribution is given by a "probability mass" function $\omega\colon X \to [0, 1] \subseteq \mathbb{R}$ with finite support $supp(\omega) = \{x \in X \mid \omega(x) \neq 0\}$ and with $\sum_x \omega(x) = 1$.
2. Each non-empty multiset $\varphi = \sum_i n_i|x_i\rangle \in \mathcal{M}(X)$ of size $n = \sum_i n_i = \|\varphi\| > 0$ gives rise to a "fractional" distribution $flrn(\varphi) \in \mathcal{D}(X)$, given by:

$$flrn\Big(\textstyle\sum_i n_i|x_i\rangle\Big) := \sum_i \frac{n_i}{n}|x_i\rangle \text{ i.e. } flrn(\varphi) := \sum_{x\in X} \frac{\varphi(x)}{\|\varphi\|}|x\rangle.$$

3. We write $\mathcal{D}(X)$ for the set of distributions on X. This $\mathcal{D}$ is functorial, like $\mathcal{M}$, see Definition 1 (3): for a function $f\colon X \to Y$ we write $\mathcal{D}(f)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$ for the function that produces the image distribution, as:

$$\mathcal{D}(f)\Big(\textstyle\sum_i r_i|x_i\rangle\Big) := \sum_i r_i|f(x_i)\rangle. \tag{6}$$

We apply the same conventions to distributions as formal sums of ket's as for multisets, so that terms $0|x\rangle$ may be ommitted, *etc.*

Such fractional distributions *flrn*(φ) have fractions as probabilities. We recall that the subset $\mathbb{Q}\!\!\rightarrow\!\!\mathbb{R}$ is dense: each real number can be expressed as limit of fractions. An analogous situation applies to distributions. In order to formulate it we use the following *total variation distance* on distributions. It is a special case of the Kantorovich-Wasserstein distance (Kantorovich and Rubinshtein, 1958). For two distributions $\omega, \omega' \in \mathcal{D}(X)$ one defines the distance $d(\omega, \omega') \in [0, 1]$ as:

$$d(\omega, \omega') \;:=\; \tfrac{1}{2} \sum_{x \in X} \big|\, \omega(x) - \omega'(x) \,\big|. \tag{7}$$

We can now formulate some basic topological properties of distributions. Stated informally, all distributions come from multisets.

Theorem 1. For a finite set X, the set $\mathcal{D}(X)$ of distributions on X, with total variation distance (Equation 7), is a compact complete metric space, containing a countable dense subset of fractional distributions, given as image of frequentist learning *flrn*, from non-empty multisets to distributions.

There is another topic that we need to introduce, namely random variables, together with their expected value—described here as validity.

Definition 3. Let X be an arbitrary set.

1. An *observable* on X is a function $p \colon X \to \mathbb{R}$. These observables are closed under pointwise sum $+$ and multiplication &. There are the always-zero and always-one observables $\mathbf{0}, \mathbf{1} \colon X \to \mathbb{R}$.

   We write $Obs(X) := \mathbb{R}^X$ for the vector space of observables on X.
2. A *random variable* is a pair (ω, p) of a distribution $\omega \in \mathcal{D}(X)$ and an observable $p \colon X \to \mathbb{R}$, on the same set X.
3. For a random variable $\big(\omega \in \mathcal{D}(X), p \colon X \to \mathbb{R}\big)$ the *expected value* is written as validity $\omega \models p$ and defined as:

$$\omega \models p \;:=\; \sum_{x \in X} \omega(x) \cdot p(x); \tag{8}$$

The expected value is commonly written as $\mathbb{E}(p)$ with the distribution ω left implicit. This may be inconvenient and confusing, especially when the distribution at hand may change, for instance in a computational setting. The notation $\mathop{\mathbb{E}}_{x \leftarrow \omega} p(x)$ fares better since it makes the distribution ω explicit, but it introduces an additional bound variable, namely the x that is sampled from ω. However, since the actual probability $\omega(x)$ of the sampled element x does not occur in this expression $\mathop{\mathbb{E}}_{x \leftarrow \omega} p(x)$, it can not be used for calculations. Hence we prefer the (new) validity notation $\omega \models p \;=\; \sum_x \omega(x) \cdot p(x)$ for the expected value of observable p in distribution ω.

4 Bayesian updating

There are two main schools in statistics, one of "frequentist" nature, assigning probabilities to data, and one of "Bayesian" kind, where probabilities are associated with hypotheses. The frequentistist approach is captured, in the discrete case, by distributions $\sum_i r_i |\, x_i\rangle \in \mathcal{D}(X)$, with probabilities r_i associated with elements / objects / data item $x_i \in X$. The Bayesian approach may be formalized in terms of belief functions $Obs(X) \to \mathbb{R}$, assigning values to observables / predicates / hypotheses / evidence. In Bayesian approaches these assignments are often called subjective, resulting from individual choices about how much value (like money) people wish to put on which possible outcomes.

The Appendix explores a mathematical perspective and describes an isomorphism (on the left in Equation 13, Appendix) that connects these Bayesian and frequentist approaches via a duality isomorphism $Hom\big(Obs(X), \mathbb{R}\big) \;\cong\; \mathcal{D}(X)$ between belief functions and distributions. Thus, one could say, the matter is solved, there is mathematically no difference between the two approaches—up to isomorphism.

Conditional probabilities are typically developed on the Bayesian side, in terms of adapted belief functions. Using the approach of the Appendix, this approach can be pulled across the duality isomorphism, to the frequentist side. It leads to a form of updating in terms of adapted distributions $\omega|_p$, see Section B in Appendix for the mathematical details. The next definition already formulates Bayesian updating of distributions with observables, in concrete form. It has been developed and used in a series of papers (Jacobs and Zanasi, 2016, 2017; Jacobs, 2017a; Cho and Jacobs, 2019; Jacobs, 2019, 2021, 2024) aimed at systematizing probabilistic updating. It is with this formalization of Bayesian updating that we can clearly formulate and prove commutativity of updating, see Proposition 1 below.

Definition 4. Let $\omega \in \mathcal{D}(X)$ be a distribution with a non-negative observable $p \colon X \to \mathbb{R}_{\geq 0}$ such that the validity $\omega \models p$ is non-zero. In that case we define the *Bayesian update* $\omega|_p \in \mathcal{D}(X)$ of ω with "evidence" p as the normalized product:

$$\omega|_p \;:=\; \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} \,|\, x\rangle. \tag{9}$$

This formulation of Bayesian updating comes alive in illustrations. We present an example first and then show how the above formulation $\omega|_p$ generalizes the traditional formulation $P(E \mid D)$.

Example 1. We consider a study involving four common species of birds: robin (R), crow (C), sparrow (S), and woodpecker (W). We start from the following species distribution (in a particular area), on the set $X = \{R, C, S, W\}$.

$$\begin{aligned} \sigma \;&=\; \tfrac{1}{4}|R\rangle + \tfrac{1}{3}|C\rangle + \tfrac{1}{4}|S\rangle + \tfrac{1}{6}|W\rangle \\ &\approx\; 0.25|R\rangle + 0.333|C\rangle + 0.25|S\rangle + 0.167|W\rangle. \end{aligned}$$

This is our prior distribution. Then a day of bird counting happens, resulting in a count observable $f \colon X \to \mathbb{N}$ with numbers:

$$f(R) \;=\; 200 \quad f(C) \;=\; 150 \quad f(S) \;=\; 50 \quad f(W) \;=\; 100.$$

We see that this observable does not really match the prior, for instance since the number of observed robins is higher than the number of crows, whereas the robin probability in σ is lower

than the crow probability. Also, the number of observed sparrows is low with respect to the woodpecker number. Hence we expect that updating σ with f will lead to a considerable change of (relative) probabilities.

We first calculate the expected value, as validity:

$$\begin{aligned}\sigma \models f &\overset{(8)}{=} \sum_{x\in X} \sigma(x)\cdot f(x)\\ &= \sigma(R)\cdot f(R) + \sigma(C)\cdot f(C) + \sigma(S)\cdot f(S) + \sigma(W)\cdot f(W)\\ &= \tfrac{1}{4}\cdot 200 + \tfrac{1}{3}\cdot 150 + \tfrac{1}{4}\cdot 50 + \tfrac{1}{6}\cdot 100\\ &= \tfrac{775}{6}.\end{aligned}$$

The updated, posterior distribution can now be computed, with this validity as normalization factor:

$$\begin{aligned}\sigma|_f &\overset{(9)}{=} \sum_{x\in X} \frac{\sigma(x)\cdot f(x)}{\sigma \models f}|x\rangle\\ &= \frac{1/4\cdot 200}{775/6}|R\rangle + \frac{1/3\cdot 150}{775/6}|C\rangle + \frac{1/4\cdot 50}{775/6}|S\rangle + \frac{1/6\cdot 100}{775/6}|W\rangle\\ &= \frac{12}{31}|R\rangle + \frac{12}{31}|C\rangle + \frac{3}{31}|S\rangle + \frac{4}{31}|W\rangle\\ &\approx 0.387|R\rangle + 0.387|C\rangle + 0.0968|S\rangle + 0.129|W\rangle.\end{aligned}$$

This posterior distribution $\sigma|_f$ incorporates the evidence of the observable f. Its robin and crow probabilities are equal, and its woodpecker probability is higher than the sparrow probability, reflecting the count outcome.

A year later a new bird count happens, resulting in a new observable $g\colon X \to \mathbb{N}$, say with $g(R) = 100$, $g(C) = 150$, $g(S) = 100$, and $g(W) = 50$. This count is more in line with the prior. One can then update $\sigma|_f$ once again, now with observable g—last year's posterior is this year's prior. The resulting second update takes the form:

$$\begin{aligned}\sigma|_f|_g &= \tfrac{12}{35}|R\rangle + \tfrac{18}{35}|C\rangle + \tfrac{3}{35}|S\rangle + \tfrac{2}{35}|W\rangle\\ &\approx 0.343|R\rangle + 0.514|C\rangle + 0.0857|S\rangle + 0.0571|W\rangle.\end{aligned}$$

This second update brings us a bit closer to the original prior σ.

Interestingly, this double update $\sigma|_f|_g$ is equal to the update $\sigma|_g|_f$ with swapped observables f, g. Thus, eventhough there is a clear order in the yearly bird counting, the mathematics of Bayesian updating ignores this order and produces the same outcome for both orders (f, g) and (g, f) of observables.

We now show how the conditional probability in traditional form fits into our form of Bayesian updating (9).

Lemma 1. Let $\omega \in \mathcal{D}(X)$ be a distribution, with a subset (event) $E \subseteq X$. We write $\mathbf{1}_E\colon X \to \mathbb{R}$ for the observable given by the indicator function of E, with associated validity:

$$\mathbf{1}_E(x) := \begin{cases} 1 & \text{if } x \in E\\ 0 & \text{if } x \notin E\end{cases} \qquad \text{and } P(E) := \omega \models \mathbf{1}_E = \sum_{x\in E} \omega(x).$$

For another subset $D \subseteq X$ one has:

1. $\mathbf{1}_E \mathbin{\&} \mathbf{1}_D = \mathbf{1}_{E\cap D}$;
2. The conditional probability $P(E \mid D)$ is obtained as validity of $\mathbf{1}_E$ in the distribution ω updated with $\mathbf{1}_D$, that is, as:

$$\omega|_{\mathbf{1}_D} \models \mathbf{1}_E = \frac{\omega \models \mathbf{1}_E \mathbin{\&} \mathbf{1}_D}{\omega \models \mathbf{1}_D} = \frac{P(E\cap D)}{P(D)} =: P(E \mid D).$$

This last equation defines the conditional probability $P(E \mid D)$.

Proof. 1. For $x \in X$, using that & is given by pointwise multiplication:

$$\begin{aligned}\big(\mathbf{1}_E \mathbin{\&} \mathbf{1}_D\big)(x) = 1 &\iff \mathbf{1}_E(x)\cdot\mathbf{1}_D(x) = 1\\ &\iff \mathbf{1}_E(x) = 1 \text{ and } \mathbf{1}_D(x) = 1\\ &\iff x \in E \text{ and } x \in D\\ &\iff x \in E\cap D \iff \mathbf{1}_{E\cap D}(x) = 1.\end{aligned}$$

2. We only have to prove the first equation, since the second one follows from the previous item. Thus:

$$\begin{aligned}\omega|_{\mathbf{1}_D} \models \mathbf{1}_E &\overset{(9)}{=} \sum_{x\in X} \omega|_{\mathbf{1}_D}(x)\cdot\mathbf{1}_E(x)\\ &\overset{(9)}{=} \sum_{x\in X} \frac{\omega(x)\cdot\mathbf{1}_D(x)}{\omega\models\mathbf{1}_D}\cdot\mathbf{1}_E(x)\\ &= \frac{\sum_{x\in X}\omega(x)\cdot(\mathbf{1}_D \mathbin{\&} \mathbf{1}_E)(x)}{P(D)}\\ &= \frac{P(E\cap D)}{P(D)}.\end{aligned}$$

□

The traditional $P(-)$ notation leaves the distribution implicit, which has many disadvantages. Most relevant in this context is that this $P(-)$ notation makes it impossible to express the commutativity of Bayesian updating, as formulated in Proposition 1 below.

We include another illustration that combines several of the topics that we discussed earlier: multisets, functoriality (for marginalization), and updating.

Example 2. Consider the following situation and questions, describing a typical update situation with observations about offspring.[1]

A friend of mine has three children aged 4 and 5 with one twin.

(a) What is the probability that there are three girls, assuming that the probability of a girl is $\frac{1}{2}$?
(b) I ring this friend's doorbell and I hear a girl's voice say that she will open the door soon. If I can assume that this is one of the three children, what is the probability that my friend has three daughters?
(c) A 4-year-old boy opens the door. Still assuming this is one of the children, what is the probability that there are three boys?

Questions (b) and (c) are independent.

We address this situation in terms of multisets of children, as in Equation 3. The situation is more complicated now since we have to use a set $C = \{B, G\}$ for the (sex of the) children but also a set $A = \{4, 5\}$ for their ages. The possible offspring configurations are (certain) multisets of size 3 over the product set $C \times A$. After a

1 See the special page https://en.wikipedia.org/wiki/Boy_or_girl_paradox.

moment's thought we see that the prior υ is a distribution of the following form.

$$\begin{aligned}\upsilon \;=\; & \tfrac{1}{16}\Big|\,2|B,4\rangle + 1|B,5\rangle\Big\rangle + \tfrac{1}{16}\Big|\,1|B,4\rangle + 2|B,5\rangle\Big\rangle \\ +\; & \tfrac{1}{8}\Big|\,1|B,4\rangle + 1|B,5\rangle + 1|G,4\rangle\Big\rangle + \tfrac{1}{16}\Big|\,2|B,5\rangle + 1|G,4\rangle\Big\rangle \\ +\; & \tfrac{1}{16}\Big|\,1|B,5\rangle + 2|G,4\rangle\Big\rangle + \tfrac{1}{8}\Big|\,1|B,4\rangle + 1|B,5\rangle + 1|G,5\rangle\Big\rangle \\ +\; & \tfrac{1}{16}\Big|\,2|B,4\rangle + 1|G,5\rangle\Big\rangle + \tfrac{1}{8}\Big|\,1|B,4\rangle + 1|G,4\rangle + 1|G,5\rangle\Big\rangle \\ +\; & \tfrac{1}{16}\Big|\,2|G,4\rangle + 1|G,5\rangle\Big\rangle + \tfrac{1}{8}\Big|\,1|B,5\rangle + 1|G,4\rangle + 1|G,5\rangle\Big\rangle \\ +\; & \tfrac{1}{16}\Big|\,1|B,4\rangle + 2|G,5\rangle\Big\rangle + \tfrac{1}{16}\Big|\,1|G,4\rangle + 2|G,5\rangle\Big\rangle\,. \end{aligned} \tag{10}$$

This υ is a distribution over multisets, using nested kets. The outer, big kets are for the probabilities, with inside the different offspring configurations in the form of a multiset. For instance, the multisets $1|B,4\rangle + 2|G,5\rangle$ and $1|G,4\rangle + 2|G,5\rangle$ in the last line capture the situations with one boy (or girl) of 4 and two girls of 5 years old.[2]

We shall write $S := \mathit{supp}\big(\upsilon\big)$ for the support of this distribution υ. This set S contains all of the 12 different multisets φ inside the big kets in Equation 10.

For the first question (a) we ask ourselves more generally what the children distribution is in this situation. It can be obtained by discarding the ages, via the first marginal of the multisets inside the big kets. This involves applying the marginalization function $\mathcal{M}(\pi_1)\colon \mathcal{M}(C \times A) \to \mathcal{M}(C)$ to these multisets, see Definition 1 3. Since we wish to apply this marginalization function $\mathcal{M}(\pi_1)$ inside the bigkets, we have to use functoriality of $\mathcal{D}$ as well, see Definition 2 3. Thus, the distribution of children marginals is obtained as:

$$\begin{aligned}\mathcal{D}\Big(\mathcal{M}(\pi_1)\Big)(\upsilon) &\overset{(6)}{=} \sum_{\varphi\in S} \upsilon(\varphi)\,\Big|\,\mathcal{M}(\pi_1)(\varphi)\Big\rangle \overset{(4)}{=} \sum_{\varphi\in S} \upsilon(\varphi)\,\Big|\,\textstyle\sum_{x,y}\varphi(x,y)|x\rangle\Big\rangle \\ &= \tfrac{1}{16}\Big|\,3|B\rangle\Big\rangle + \tfrac{1}{16}\Big|\,3|B\rangle\Big\rangle + \tfrac{1}{8}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle \\ &\quad + \tfrac{1}{16}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle + \tfrac{1}{16}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle \\ &\quad + \tfrac{1}{8}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle + \tfrac{1}{16}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle \\ &\quad + \tfrac{1}{8}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle + \tfrac{1}{16}\Big|\,3|G\rangle\Big\rangle + \tfrac{1}{8}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle \\ &\quad + \tfrac{1}{16}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle + \tfrac{1}{16}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle \\ &= \tfrac{1}{8}\Big|\,3|B\rangle\Big\rangle + \tfrac{3}{8}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle \\ &\quad + \tfrac{3}{8}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle + \tfrac{1}{8}\Big|\,3|G\rangle\Big\rangle\,. \end{aligned}$$

The answer to question (a) is thus $\frac{1}{8}$, the probability associated in the last line with the three girls multiset $3|G\rangle$.

2 One can obtain the distribution υ in Equation 10 itself via conditioning, namely from the multinomial distribution, of draws of size three from the uniform distribution on $C \times A$. One updates this multinomial distribution by keeping only those multiset φ in which both ages occur, that is, for which the support $\mathit{supp}(\mathcal{M}(\pi_2)(\varphi)) \subseteq A$ of the second marginal of φ has two elements. This construction of υ distracts from the main line, so we decided to simply present the relevant prior distribution in Equation 10.

The interested reader may wish to check that taking the second marginals yields the (expected) age distribution of the form:

$$\mathcal{D}\Big(\mathcal{M}(\pi_2)\Big)(\upsilon) \;=\; \tfrac{1}{2}\Big|\,2|4\rangle + 1|5\rangle\Big\rangle + \tfrac{1}{2}\Big|\,1|4\rangle + 2|5\rangle\Big\rangle\,.$$

For question (b) we define an observable $g\colon S \to \{0,1\}$ which is 1 if and only if there is at least one girl:

$$g(\varphi) = 1 \iff \mathcal{M}(\pi_1)(\varphi)(g) \geq 1 \iff \varphi(g,4) + \varphi(g,5) \geq 1.$$

This observable g is $\{0,1\}$-valued and may be identified with a subset of S, as in Lemma 1. Updating υ with g involves removing the multisets $\varphi \in S$ with $g(\varphi) = 0$, that is, with boys only, and then renormalising. The normalization factor is the validity:

$$\upsilon \models g \;=\; \sum_{\varphi\in S} \upsilon(\varphi)\cdot g(\varphi) \;=\; \tfrac{7}{8}$$

The answer to question (b) is obtained by computing the update $\upsilon|_g$ and taking its children marginal, as before. This yields:

$$\begin{aligned}&\mathcal{D}\Big(\mathcal{M}(\pi_1)\Big)\big(\upsilon|_g\big) \\ &= \mathcal{D}\Big(\mathcal{M}(\pi_1)\Big)\Big(\tfrac{1}{7}\Big|\,1|B,4\rangle + 1|B,5\rangle + 1|G,4\rangle\Big\rangle \\ &\qquad + \tfrac{1}{14}\Big|\,2|B,5\rangle + 1|G,4\rangle\Big\rangle + \tfrac{1}{14}\Big|\,1|B,5\rangle + 2|G,4\rangle\Big\rangle \\ &\qquad + \tfrac{1}{14}\Big|\,2|B,4\rangle + 1|G,5\rangle\Big\rangle + \tfrac{1}{7}\Big|\,1|B,4\rangle + 1|B,5\rangle + 1|G,5\rangle\Big\rangle \\ &\qquad + \tfrac{1}{7}\Big|\,1|B,4\rangle + 1|G,4\rangle + 1|G,5\rangle\Big\rangle + \tfrac{1}{7}\Big|\,1|B,5\rangle + 1|G,4\rangle 1|G,5\rangle\Big\rangle \\ &\qquad + \tfrac{1}{14}\Big|\,2|G,4\rangle + 1|G,5\rangle\Big\rangle + \tfrac{1}{14}\Big|\,1|B,4\rangle + 2|G,5\rangle\Big\rangle \\ &\qquad + \tfrac{1}{14}\Big|\,1|G,4\rangle + 2|G,5\rangle\Big\rangle\Big) \\ &= \tfrac{3}{7}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle + \tfrac{3}{7}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle + \tfrac{1}{7}\Big|\,3|G\rangle\Big\rangle\,. \end{aligned}$$

We can conclude that after seeing one girl the probability that there are three girls has risen from $\frac{1}{8}$ to $\frac{1}{7}$. As an aside: the distribution of age marginals remains the same after this update.

What happens when we see a 4-year old boy? We capture this via an event / observable $b_4\colon S \to \{0,1\}$ with $b_4(\varphi) = 1$ iff $\varphi(B,4) \geq 1$. Its validity $\upsilon \models b_4$ in the prior distribution υ is $\frac{7}{12}$. We leave it to the interested reader to verify that the distributions of children / age marginals, after update with b_4, are:

$$\begin{aligned}\mathcal{D}\Big(\mathcal{M}(\pi_1)\Big)\big(\upsilon|_{b_4}\big) \;=\; & \tfrac{1}{5}\Big|\,3|B\rangle\Big\rangle + \tfrac{1}{10}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle \\ & + \tfrac{3}{10}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle \\ \mathcal{D}\Big(\mathcal{M}(\pi_2)\Big)\big(\upsilon|_{b_4}\big) \;=\; & \tfrac{3}{5}\Big|\,2|4\rangle + 1|5\rangle\Big\rangle + \tfrac{2}{5}\Big|\,1|4\rangle + 2|5\rangle\Big\rangle\,. \end{aligned}$$

The first equation answers question (c): the probability of three boys is $\frac{1}{5}$, having seen one 4-year old boy. It is higher than the probability of seeing three girls, given that there is at least one girl? The boy-of-4 observation excludes more cases and the remaining cases thus get higher probability, after re-normalization.

The second equation about the age marginals shows that the configuration with two 4-year olds is more likely, after seeing at least one 4-year old (boy). This makes sense.

We can still ask what we can infer if we have seen both a girl and a boy-of-4. As before the order of updating is irrelevant:

$\upsilon|_g|_{b_4} = \upsilon|_{b_4}|_g$. In that situation there are 5 multisets left, out of the original 12, in υ in Equation 10. The distributions of marginals are:

$$\mathcal{D}\Big(\mathcal{M}(\pi_1)\Big)\big(\upsilon|_g|_{b_4}\big) = \tfrac{5}{8}\Big|\,2|B\rangle + 1|G\rangle\Big\rangle + \tfrac{3}{8}\Big|\,1|B\rangle + 2|G\rangle\Big\rangle$$
$$\mathcal{D}\Big(\mathcal{M}(\pi_2)\Big)\big(\upsilon|_g|_{b_4}\big) = \tfrac{5}{8}\Big|\,2|4\rangle + 1|5\rangle\Big\rangle + \tfrac{3}{8}\Big|\,1|4\rangle + 2|5\rangle\Big\rangle.$$

We conclude by proving in general what we have already seen several times, namely that multiple Bayesian updates commute (Equation 1). We do so by using the conjunction $p \;\&\; q$ (pointwise multiplication) of observables, in order to emphasize the close connection between commutativity of conjunction and of updating. The result below already occurs in Jacobs (2019, Lem. 4.1), together with a generalized formulation of Bayes' rule for observables. A proof is included for completeness.

Proposition 1. Let $\omega \in \mathcal{D}(X)$ be a distribution with two non-negative observables $p, q\colon X \to \mathbb{R}_{\geq 0}$. Then, assuming that the relevant validities are non-zero,

$$\omega|_p|_q = \omega|_{p\&q} = \omega|_{q\&p} = \omega|_q|_p.$$

Proof. We only have to prove the first equation, since the commutativity of & is obvious (multiplication of numbers is commutative) and the last equation is an instance of the first (with p, q swapped). Using the functional description for distributions, we have for $x \in X$,

$$\begin{aligned}\omega|_p|_q(x) &\overset{(9)}{=} \frac{\omega|_p(x)\cdot q(x)}{\omega|_p \models q} \\ &\overset{(9)}{=} \frac{\frac{\omega(x)\cdot p(x)}{\omega\models p}\cdot q(x)}{\sum_y \frac{\omega(y)\cdot p(y)}{\omega\models p}\cdot q(y)} \\ &= \frac{\omega(x)\cdot(p\;\&\;q)(x)}{\omega\models p\;\&\;q} \overset{(9)}{=} \omega|_{p\&q}(x).\end{aligned}$$ □

5 Concluding remarks

The Bayesian approach is popular in cognition theory, where the human mind is seen as a Bayesian prediction and inference engine, see for instance the recent books (Griffiths et al., 2024; Parr et al., 2022). In that line of work the mismatch caused by the commutativity of Bayesian updating does not get much attention. It is however known in the literature, see notably (Uzan, 2023). One way out is to switch from classical to quantum probability, where conjunction and updating are non-commutative. This has led to a new line of "quantum" cognition theory (see e.g. Busemeyer and Bruza, 2012; Yearsley and Busemeyer, 2016; or Jacobs, 2017b which is similar in style to this article).

When we take the commutativity of Bayesian updating seriously, the proper data structure to deal with multiple updates is: a multiset of observables. Indeed, as we have seen in Section 2, multisets abstract from lists by ignoring the order. This perspective is elaborated in Jacobs (2024), where the different update mechanisms of Pearl and Jeffrey (Jacobs, 2019), and also the variational free update mechanism from predictive coding (Friston, 2009; Tull et al., 2023), are formulated in terms of such multisets of observables. Jeffrey's rule is non-commutative, but in a special way, namely for multiple such (non-singleton) multisets. All this suggests that the topic of commutativity may be a decisive element in further developing probabilistic perspectives in cognition and in AI.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

BJ: Writing – original draft, Methodology, Formal analysis, Investigation, Conceptualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcogn.2025.1623227/full#supplementary-material

References

Alfsen, E. (1971). *Compact Convex Sets and Boundary Integrals, Volume 57 of Ergebnisse der Mathematik und ihrer Grenzgebiete*. Cham: Springer.

Awodey, S. (2006). *Category Theory. Oxford Logic Guides*. Oxford: Oxford University Press.

Busemeyer, J., and Bruza, P. (2012). *Quantum Models of Cognition and Decision*. Cambridge: Cambridge University Press.

Chaput, P., Danos, V., Panangaden, P., and Plotkin, G. (2014). Approximating Markov processes by averaging. *J. ACM*, 61, 1–45. doi: 10.1145/2537948

Cheng, E. (2022). *The Joy of Abstraction. An Exploration of Math, Category Theory, and Life*. Cambridge: Cambridge University Press.

Cho, K., and Jacobs, B. (2019). Disintegration and Bayesian inversion via string diagrams. *Math. Struct. Comp. Sci.* 29, 938–971. doi: 10.1017/S0960129518000488

Conway, J. (1990). *A Course in Functional Analysis. Graduate Texts in Mathematics 96*, 2nd Edn. Cham: Springer.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci*. 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Fritz, T. (2020). A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Adv. Math*. 370:107239. doi: 10.1016/j.aim.2020.107239

Gentili, P. (2021). Establishing a new link between fuzzy logic, neuroscience, and quantum mechanics through Bayesian probability: perspectives in artificial intelligence and unconventional computing. *Molecules* 26:5987. doi: 10.3390/molecules26195987

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev*. 102, 684–704.

Griffiths, T., Chater, N., and Tenenbaum, J. (2024). *Bayesian Models of Cognition. Reverse Engineering the Mind*. Cambridge, MA: MIT Press.

Jacobs, B. (2017a). Hyper normalisation and conditioning for discrete probability distributions. *Log. Methods Comp. Sci*. 13. doi: 10.23638/LMCS-13(3:17)2017

Jacobs, B. (2017b). Quantum effect logic in cognition. *J. Math. Psychol*. 81, 1–10. doi: 10.1016/j.jmp.2017.08.004

Jacobs, B. (2017c). A recipe for state and effect triangles. *Log. Methods Comp. Sci*. 13, 1–26. doi: 10.23638/LMCS-13(2:6)2017

Jacobs, B. (2019). The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. *J. Artif. Intell. Res*. 65, 783–806. doi: 10.1613/jair.1.11349

Jacobs, B. (2021). "Learning from what's right and learning from what's wrong," in *Mathematical Foundations of Programming Semantics, number 351 in Elect. Proc. in Theor. Comp. Sci*., ed. A. Sokolova, 116–133. doi: 10.4204/EPTCS.351.8

Jacobs, B. (2022). Urns & tubes. *Compositionality* 4. doi: 10.32408/compositionality-4-4

Jacobs, B. (2024). Getting wiser from multiple data: probabilistic updating according to Jeffrey and Pearl. *arXiv* [Preprint] arXiv:2405.12700. doi: 10.48550/arXiv.2405.12700

Jacobs, B. (2025). *Structured Probabilitistic Reasoning*. Available online at: http://www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf (Accessed June 2025).

Jacobs, B., Mandemaker, J., and Furber, R. (2016). The expectation monad in quantum foundations. *Inf. Comput*. 250, 87–114. doi: 10.1016/j.ic.2016.02.009

Jacobs, B., and Zanasi, F. (2016). "A predicate/state transformer semantics for Bayesian learning," in *Math. Found. of Programming Semantics, number 325 in Elect. Notes in Theor. Comp. Sci*., ed. L. Birkedal (Amsterdam: Elsevier), 185–200.

Jacobs, B., and Zanasi, F. (2017). "A formal semantics of influence in Bayesian reasoning," in *Math. Found. of Computer Science, Volume 83 of LIPIcs*, eds. K. Larsen, H. Bodlaender, and J.-F. Raskin (Wadern: Schloss Dagstuhl), 21:1–21:14.

Johnson, N., and Kotz, S. (1977). *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. Hoboken, NJ: John Wiley.

Johnstone, P. (1982). *Stone Spaces. Number 3 in Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.

Kantorovich, L., and Rubinshtein, G. (1958). On a space of totally additive functions. *Vestnik Leningrad Univ*. 13, 52–59.

Leinster, T. (2014). *Basic Category Theory. Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press. doi: 10.48550/arXiv.1612.09375

Mahmoud, H. (2008). *Pólya Urn Models*. Boca Raton, FL: Chapman and Hall.

Panangaden, P. (2009). *Labelled Markov Processes*. London: Imperial College Press.

Parr, T., Pezzulo, G., and Friston, K. (2022). *Active Inference. The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA: MIT Press.

Perrone, P. (2024). *Starting Category Theory*. Singapore: World Scientific.

Pierce, B. (1991). *Basic Category Theory for Computer Scientists*. Cambridge, MA: MIT Press.

Ross, S. (2018). *A First Course in Probability*, 10th Edn. Upper Saddle River, NJ: Pearson Education.

Simons, H. (2011). *An Introduction to Category Theory*. Cambridge: Cambridge University Press.

Tull, S., Kleiner, J., and Smithe, T. S. C. (2023). Active inference in string diagrams: a categorical account of predictive processing and free energy. *arXiv* [Preprint] arXiv.2308.00861. doi: 10.48550/arXiv.2308.00861

Uzan, P. (2023). Bayesian rationality revisited: integrating order effects. *Found. Sci*. 28, 507–528. doi: 10.1007/s10699-022-09838-0

Yearsley, J., and Busemeyer, J. (2016). Quantum cognition and decision theories: a tutorial. *J. Math. Psychol*. 74, 99–116.