entropy

MDPI

Article

A Formal Framework for Knowledge Acquisition: Going beyond Machine Learning

Ola Hössjer [1], Daniel Andrés Díaz-Pachón [2,*] and J. Sunil Rao [2]

[1] Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden
[2] Division of Biostatistics, University of Miami, Miami, FL 33136, USA
* Correspondence: ddiaz3@miami.edu

Abstract: Philosophers frequently define knowledge as justified, true belief. We built a mathematical framework that makes it possible to define learning (increasing number of true beliefs) and knowledge of an agent in precise ways, by phrasing belief in terms of epistemic probabilities, defined from Bayes' rule. The degree of true belief is quantified by means of active information I^+: a comparison between the degree of belief of the agent and a completely ignorant person. Learning has occurred when either the agent's strength of belief in a true proposition has increased in comparison with the ignorant person ($I^+ > 0$), or the strength of belief in a false proposition has decreased ($I^+ < 0$). Knowledge additionally requires that learning occurs for the right reason, and in this context we introduce a framework of parallel worlds that correspond to parameters of a statistical model. This makes it possible to interpret learning as a hypothesis test for such a model, whereas knowledge acquisition additionally requires estimation of a true world parameter. Our framework of learning and knowledge acquisition is a hybrid between frequentism and Bayesianism. It can be generalized to a sequential setting, where information and data are updated over time. The theory is illustrated using examples of coin tossing, historical and future events, replication of studies, and causal inference. It can also be used to pinpoint shortcomings of machine learning, where typically learning rather than knowledge acquisition is in focus.

Keywords: active information; Bayes' rule; counterfactuals; epistemic probability; learning, justified true belief; knowledge acquisition; replication studies

Citation: Hössjer, O.; Díaz-Pachón, D.A.; Rao, J.S. A Formal Framework for Knowledge Acquisition: Going beyond Machine Learning. *Entropy* **2022**, *24*, 1469. https://doi.org/10.3390/e24101469

Academic Editor: Deniz Gençağa

Received: 3 September 2022
Accepted: 7 October 2022
Published: 14 October 2022

1. Introduction

1.1. The Present Article

The process by which cognitive agents acquire knowledge is complicated, and has been studied from different perspectives within educational science, psychology, neuroscience, cognitive science, and social science [1]. Philosophers usually distinguish between three types of knowledge [2]: acquaintance knowledge (to get to know other persons), knowledge how (to learn certain skills), and knowledge that (to learn about propositions or facts). Mathematically, acquaintance knowledge has been studied via trees and networks, for instance, in small-world-type models and rumor-spreading models [3–5]. Knowledge how has been widely developed in education and psychology, since the middle of the twentieth century, by means of testing and psychometry, using classical statistics [6–8].

The purpose of this paper is to formulate knowledge that in mathematical terms. Our starting point is to define knowledge that as justified true belief (JTB), which generally is agreed to constitute at least a sufficient condition for such knowledge [9,10]. The primary tools will be the concepts of truth, probabilities, and information theory. Probabilities, in addition to logic, are used to formulate mechanisms of reasoning in order to define beliefs [11,12]. More specifically, a Bayesian approach with subjective probabilities will be used to quantify rational agents' degrees of beliefs in a proposition. These subjective probabilities may vary between agents, but since each agent is assumed to be rational,

its probabilities satisfy basic axioms of probability [13]. This is also referred to as the personalistic view of probabilities in [14].

The degree of belief in a proposition is associated with some type of randomness or uncertainty regarding the truth of the proposition. It is helpful in this context to distinguish between ontological randomness (genuine randomness regarding the truth of the proposition) and epistemic randomness (incomplete knowledge about propositions that are either true or false). Here the focus will be on epistemic randomness, and following [15], subjective probabilities are referred to as epistemic probabilities. The epistemic randomness assumption that each proposition has a fixed truth value can be viewed as a frequentist component of our framework.

To use epistemic probabilities in a wider context of knowledge that (subsequently simply referred to as knowledge), we incorporate degrees of beliefs within a framework of parallel worlds in order to define more clearly what JTB means. These parallel worlds correspond to parameters of a statistical model and a second frequentist notion of one parameter being true, whereas the others are counterfactuals [16]. An agent's maximal possible discernment between worlds is described in terms of the σ-algebra $\mathcal{G}$. The agent's degrees of belief are obtained through Bayes' rule from prior belief and data [17], in such a way that it is not possible to discern between worlds beyond the limits set by $\mathcal{G}$.

Learning is associated with increased degrees of true belief, although these beliefs need not necessarily be justified. More specifically, the agent's degree of belief in a proposition is compared to that of an ignorant person. This corresponds to an hypothesis test within a frequentist framework. More specifically, the null hypothesis of a proposition being true is tested against an alternative hypothesis that the proposition is false. As a test statistic, we use active information I^+ [18–20], which quantifies how much the agent has learned about the truth value of the proposition compared to an ignorant person. In particular, learning has occurred when the agent's degree of belief in a true proposition is larger than that of an ignorant person ($I^+ > 0$), or if the agent's degree of belief in a false proposition is less than that of an ignorant person ($I^+ < 0$). In either case, $\mathcal{G}$ sets a limit in terms of the maximal amount of possible learning. Learning is, however, not sufficient for knowledge acquisition, since the latter concept also requires that the true belief is justified, or has been formed for the right reason. Knowledge acquisition is defined as a learning process where the agent's degree of belief in the true world is increased, corresponding to a more accurate estimate of the true world parameter. Thus, knowledge acquisition goes beyond learning in that it also deals with the "justified" part of the JTB condition. It is related to consistency of a posterior distribution, a notion that is meaningful only within our hybrid frequentist/Bayesian approach.

To the best of our knowledge, the hybrid frequentist/Bayesian approach has only been used in the context of Bayesian asymptotic theory (Section 7.2), but not as a general tool for modeling the distinction between learning and knowledge acquisition. Although the concept of a true world (or the true state of affairs) is used in the context of Bayesian decision theory and its extensions, such as robust Bayesian inference and belief functions based on the Dempster–Shafer theory [21–24], the goal is then to maximize an expected utility (or to minimize an expected cost) of the agent that makes the decision. In our context, the Bayesian approach is only used to formulate beliefs as posterior distributions, whereas the criteria for learning (probabilities of rejecting a false or true proposition) and knowledge acquisition (consistency) are frequentist. Given that a model with one unique, true world is correct, the frequentist error probability and consistency criteria are objective, since they depend on the true world. No such criteria exist within a purely Bayesian framework.

Illustration 1. *In order to illustrate our approach for modeling learning and knowledge acquisition, we present an example that will be revisited several times later on. A teacher (the agent) wants to evaluate whether a child has learned addition. The teacher gives the student a home assignment test with two-choice answers, one right and one wrong, to measure the proposition S: "The child is expected to score well on the test." In this case, we have a set* $\mathcal{X} = \{x_1, x_2, x_3\}$ *of three possible*

worlds. An ignorant person who does not ask for help is expected to have half her questions right and half her questions wrong (x_1). A child who knows addition is expected to get a large fraction of the answers right (x_2). However, there is also a third alternative, where an ignorant student asks for help and is expected to have a high score for that reason (x_3). Notice in particular that S is true only for the two worlds of the set $A = \{x_2, x_3\}$. If the child answers substantially more questions right than wrong, the active information will be positive and the teacher learns S. However, this learning that S is true does not represent knowledge of whether the student knows how to add, since the teacher is not able to distinguish x_2 from x_3. Now, let us say that the test has only two questions. In this setting, it is expected that an ignorant person has one question right and one wrong. However, it is also highly probable that even if the child does not know his sums well, he can answer the two questions in the right way. In this case, the teacher has not learned substantially about S (nor attained knowledge of whether the student knows how to add). The reason is that, since the test has only two questions, the teacher cannot exclude any of x_1, x_2, and x_3. The more questions the test has, and if the student scores well, the more certain the teacher is that either x_2 or x_3 is true, that is, the more he learns about S. If the student is also monitored during the exam, alternative x_3 is excluded and the teacher knows that x_2 is true; that is, the teacher not only learns about S, but also acquires knowledge that the student knows how to add.

Each of the following sections contains remarks and illustrations like the previous one. At the end of the paper, a whole section with multiple examples will explore deeper how the model works in practice.

1.2. Related Work

Other contributions have been made to developing a mathematical framework for learning and knowledge acquisition. Hopkins [25] studied the theoretical properties of two different models of learning in games, namely, reinforcement learning and stochastic fictitious play. He developed an equivalence relation between the two under a variety of different scenarios with increasing degrees of structure. Stoica and Strack [26] introduced a stochastic model for acquired knowledge and showed that empirical data fit the estimated outcomes of the model well, using data from student performance in university-level classes. Taylor [27] proposed a model using the notion of concept lattices and the mathematical theory of closure spaces to describe knowledge acquisition and organization. However, none of these works has been developed through basic concepts in probability and information theory the way we do here. Our approach permits important generalizations which cover a wide range of real-life scenarios.

2. Possible Worlds, Propositions, and Discernment

Consider a collection $\mathcal{X}$ of *possible* worlds, of which $x_0 \in \mathcal{X}$ is the true world, and all other worlds $x \in \mathcal{X} \setminus \{x_0\}$ are counterfactuals. We will regard x as a statistical parameter, and the fact that this parameter has a true but unknown value x_0 corresponds to a frequentist assumption. The set $\mathcal{X}$ is the parameter space of interest, and it is assumed to be either finite or a bounded and open subset of Euclidean space $\mathbb{R}^q$ of dimension q. Let S be a proposition (or statement), and impose a second frequentist assumption that S is either true or false, although the truth value of S may depend on the world $x \in \mathcal{X}$. Define a binary-valued truth function $f : \mathcal{X} \to \{0,1\}$ by $f(x) = 1$ or 0, depending on whether S is true or not in world x. The set $A = \{x \in \mathcal{X};\, f(x) = 1\}$ consists of all worlds for which S is a true proposition. Although there is one-to-one correspondence between f and A, in the sequel it will be convenient to use both notions. The simplest truth scenario of S is one for which the truth value of S is unique for the true world, i.e.,

$$A_0 = \begin{cases} \{x_0\}, & \text{if } f(x_0) = 1, \\ \mathcal{X} \setminus \{x_0\}, & \text{if } f(x_0) = 0. \end{cases} \tag{1}$$

x_0 being unique and f being binary-valued together correspond to a framework of epistemic randomness, where the actual truth value $f(x_0)$ of S is either 0 or 1. S is referred to as

falsifiable [28] if it is logically possible (in principle) to find a data set D implying that the truth value of S is 0, or equivalently, that none of the worlds in A is true. It is possible though to falsify S without knowing x_0.

3. Probabilities

3.1. Degrees of Beliefs and Sigma Algebras

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. When $\mathcal{X}$ is finite, $\mathcal{F}$ consists of all subsets of $\mathcal{X}$ (i.e., $\mathcal{F} = 2^{\mathcal{X}}$); otherwise, $\mathcal{F}$ is the class of Borel sets. The Bayesian part of our approach is to quantify an agent's belief in which a world is true by means of an epistemic probability measure P on the measurable space $(\mathcal{X}, \mathcal{F})$, whereas the beliefs of an ignorant person follow another probability measure P_0. It is often assumed that

$$P_0(B) = \frac{|B|}{|\mathcal{X}|}, \quad \forall B \in \mathcal{F}, \tag{2}$$

is the uniform probability measure that maximizes entropy among all probability measures on $(\mathcal{X}, \mathcal{F})$, where $|\cdot|$ refers to the cardinality for finite $\mathcal{X}$ and to the Lebesgue measure for continuous $\mathcal{X}$. Then, (2) corresponds to a maximal amount of ignorance about which possible world is true [29]. Sometimes (as in Example 5 below) some general background knowledge is assumed also for the ignorant person, so that P_0 differs from (2).

The agent's and the ignorant person's strength of belief in S are quantified by $P(A)$ and $P_0(A)$, respectively. Following [15], it is helpful to interpret $P(A)$ and $P_0(A)$ as the agent's and the ignorant person's predictions of the physical probability $f(x_0) \in \{0, 1\}$ of S. Whereas P and P_0 involve epistemic uncertainty, the physical probability is an indicator for the real (physical) event that S is true or not.

When an agent's belief P is formed, it is assumed that any information accessible to him, beyond that of the ignorant person, belongs to a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$. This means that the agent has no more knowledge of how to discern events in $\mathcal{G}$ than the ignorant person, if this discernment requires that he considers events in $\mathcal{F}$ that do not belong to $\mathcal{G}$. Mathematically, this corresponds to a requirement

$$E_P[g \mid \mathcal{G}'] = E_{P_0}[g \mid \mathcal{G}'], \tag{3}$$

for all $\mathcal{F}$-measurable functions $g : \mathcal{X} \to \mathbb{R}$, and all sigma algebras $\mathcal{G}'$ such that $\mathcal{G} \subseteq \mathcal{G}' \subseteq \mathcal{F}$. It is assumed, on the left-hand side of (3), that g is a random variable defined on the probability space $(\mathcal{X}, \mathcal{F}, P)$, whereas g is defined on the probability space $(\mathcal{X}, \mathcal{F}, P_0)$ on the right-hand side of (3). It follows from (3) that $\mathcal{G}$ sets the limit in terms of the agent's possibility to form propositions about which world is true. Therefore, $\mathcal{G}$ is referred to as the agent's maximal possible *discernment* about which world is true. It follows from (3) that

$$\begin{aligned} P(A) &= E_P[f] \\ &= E_P\{E_P[f \mid \mathcal{G}]\} \\ &= E_P\{E_{P_0}[f \mid \mathcal{G}]\}. \end{aligned} \tag{4}$$

The minimal amount of discernment corresponds to the trivial σ-algebra $\mathcal{G}_0 = \{\emptyset, \mathcal{X}\}$. Whenever (3) holds with $\mathcal{G} = \mathcal{G}_0$, necessarily $P = P_0$. This corresponds to removing the outer expectation on the right-hand side of (4), so that

$$\begin{aligned} P_0(A) &= E_{P_0}[f] \\ &= E_{P_0}[f \mid \mathcal{G}_0]. \end{aligned} \tag{5}$$

Remark 1. *Suppose there exists an oracle or omniscient agent $\mathcal{O}$ that is able to discern between all possible worlds and also knows x_0. Mathematically, the discernment requirement means that $\mathcal{O}$ has knowledge about all sets in a σ-algebra $\mathcal{F}$ that corresponds to a maximal amount of discernment between possible worlds. We will assume that f is measurable with respect to $\mathcal{F}$, so that A is*

measurable (i.e., $A \in \mathcal{F}$). Knowledge of $\mathcal{F}$ is, however, not sufficient for knowing A, since A may involve x_0, as in (1). By this we mean that if the agent knows $\mathcal{F}$, and if A involves x_0, then there are several candidates of A for the agent, and he does not know a priori which one of these candidates is the actual A. However, since $\mathcal{O}$ knows $\mathcal{F}$ and x_0, he also knows A. It follows that $\mathcal{O}$ knows that S is true for all worlds in (the actual) A, and that S is false for all worlds outside of (the actual) A. That is, the oracle knows for which possible worlds the proposition S is true.

As mentioned in Remark 1, the truth function f is measurable with respect to the maximal σ-algebra $\mathcal{F}$. However, depending on how $\mathcal{G}$ is constructed, and whether A involves x_0 or not, the set A may or may not be known to the agent. Therefore, when A involves x_0, the agent may not be able to compute $P_0(A)$ and $P(A)$ himself. Although he is able to compute $P_0(B)$ and $P(B)$ for all $B \in \mathcal{F}$, since he does not know x_0, it follows that he does not know which of these sets B equals A. Therefore, he does not know $P(A)$ and $P_0(A)$, unless $P(B) = P(A)$ and $P_0(B) = P_0(B)$, respectively, for all B that are among the agent's candidates for the set A. For instance, suppose $\mathcal{X} = \{1,2,3\}$, $P(1) = 1/5$, $P(2) = P(3) = 2/5$, and $A = \{3\}$. If the agent's candidates for A are $\{1\}$, $\{2\}$, and $\{3\}$, then the agent does not know $P(A)$. On the other hand, if the agent's candidates for A are $\{2\}$ and $\{3\}$, then he knows $P(A)$, although he does not know A.

As will be seen from the examples of Section 8, it is often helpful (but not necessary) to construct $\mathcal{G}$ as the σ-algebra that is generated by a random variable Y whose domain is $\mathcal{X}$ (i.e., $\mathcal{G} = \sigma(Y)$). This means that Y determines the collection $\mathcal{G}$ of subsets of $\mathcal{X}$ for which the agent is free to form beliefs beyond that of the ignorant person. Typically, Y highlights the way in which information is lost by going from $\mathcal{F}$ to $\mathcal{G}$. For instance, suppose $\mathcal{X} = [0,\infty)$ and $Y : [0,\infty) \to \{0,1,2,\ldots\}$ is defined by $Y(x) = [x/\delta]$ for some $\delta > 0$; then, $\mathcal{G} = \sigma(\{[0,\delta),[\delta,2\delta),\ldots\})$ is the sigma-algebra obtained by from a quantization procedure with accuracy δ.

3.2. Bayes' Rule and Posterior Probabilities

A Bayesian approach will be used to define the agent's degree of belief P. To this end, we regard $x \in \mathcal{X}$ as a parameter of a statistical model and that the agent has access to data $d \in \mathcal{D}$. The agent assumes that (x,d) is an observation of a random variable $(X,D) : \Omega \to \mathcal{X} \times \mathcal{D}$ defined on some sample space Ω. The joint distribution of the parameter X and data D, according to the agent's beliefs, is $dQ(x,D) = dP_0(x)L(D|x)dD$. This is a probability measure on subsets of $\mathcal{X} \times \mathcal{D}$, with prior distribution P_0 of the parameter X, and with a conditional distribution $L(D|x) = dQ(D|x)/dD$ that corresponds to the likelihood of data D. A posterior probability

$$P(A) = Q(A \mid D) = \int_A dQ(x \mid D) \tag{6}$$

of A is formed by updating the prior distribution P_0 based on data D. It is assumed that the likelihood $x \to L(D \mid x)$ is measurable with respect to $\mathcal{G}$, so that data conform with the agent's maximal possible discernment between possible worlds. The likelihood function $x \to L(D \mid x)$ includes the agent's *interpretation* of D. Although this interpretation may involve a subjective part, it is still assumed that the agent is not willing to speculate about possible worlds beyond the limits set by $\mathcal{G}$. That is, whenever the agent discerns events in $\mathcal{G}$ beyond the limits set by $\mathcal{G}$, this discernment is the same as for an ignorant person.

Remark 2. *To account for the possibility that the agent still speculates beyond the limits set by external data, $\mathcal{G} = \sigma(\mathcal{G}_{ext}, \mathcal{G}_{int})$ could be defined as the smallest σ-algebra containing the σ-algebras $\mathcal{G}_{ext}$ and $\mathcal{G}_{int}$ that originate from external data D_{ext} and internal data D_{int} (the agent's internal experiences, such as dreams and revelations, respectively). Note, however, that $x \to L(D \mid x)$ is subjective, even when internal data are absent, since agents might interpret external data in different ways, due to the way in which they perceive such data and incorporate previous life experience.*

From Bayes' rule we find that the posterior distribution satisfies

$$\begin{aligned} P(A) &= Q(A \mid D) \\ &= \int_A dQ(x \mid D) \\ &= \frac{L(D \mid A)P_0(A)}{L(D)} \\ &= \frac{\int_A L(D \mid x)dP_0(x)}{\int_{\mathcal{X}} L(D \mid x)dP_0(x)}. \end{aligned} \tag{7}$$

A couple of additional reasons reinforce the subjectivity of P: the prior P_0 might be subjective, and acquisition of data D might vary between agents [30]. Additionally, acquisition of data D will not necessarily make P more concentrated around the true world x_0, since it is possible that the data themselves are biased or that the agent interprets the data in a sub-optimal way.

Since the likelihood function is measurable with respect to $\mathcal{G}$, it follows from (4) that the agent's belief P, after having observed D, does not lead to a different discernment between possible worlds beyond $\mathcal{G}$ than for an ignorant person. Given $\mathcal{G}$, together with an unlimited amount of unbiased data that the agent interprets correctly, the *$\mathcal{G}$-optimal choice of P* is

$$P(B) = \mathbf{1}(x_0 \in B), \quad \forall B \in \mathcal{G}. \tag{8}$$

Equations (4) and (8) uniquely define the $\mathcal{G}$-optimal choice of P. Whenever $\mathcal{G} \subset \mathcal{F}$ is a proper subset of the maximal σ-algebra $\mathcal{F}$, the measure P in (8) is not the same thing as a point mass δ_{x_0} at x_0. On the other hand, for an oracle with a maximal amount of knowledge about which world is true, $\mathcal{G} = \mathcal{F}$, (8) reduces to a point mass at the true world—i.e.,

$$P = \delta_{x_0} \iff P(B) = \mathbf{1}(x_0 \in B), \quad \forall B \in \mathcal{F}. \tag{9}$$

Remark 3. *An extreme example of biased beliefs is a true-world-excluding probability measure, with support that does not include x_0:*

$$supp(P) \subset \mathcal{X} \setminus \{x_0\}. \tag{10}$$

Another example is a correct-proposition-excluding probability measure, with support that excludes all worlds x with a correct value $f(x) = f(x_0)$ of S:

$$supp(P) \subset \begin{cases} A^c = \mathcal{X} \setminus A, & x_0 \in A, \\ A, & x_0 \notin A. \end{cases} \tag{11}$$

Illustration 2 (Continuation of Illustration 1)**.** *Suppose data $D \in \mathcal{D} = \{0, 1, \ldots, 10\}$ are available to the teacher (the agent) in terms of the number of correct answers of a home assignment test with 10 questions. The prior $P_0(x_i) = 1/3$ is uniform on $\mathcal{X} = \{x_1, x_2, x_3\}$, whereas data $D|x_i \sim Bin(10, \pi_i)$ have a binomial distribution with probabilities $\pi_1 = 0.5$ and $\pi_2 = \pi_3 = 0.8$ of answering each question correctly, for a student that either guesses or has math skills/asks for help. Let d be the observed value of D. Since data have the same likelihood ($L(d|x_2) = L(d|x_3)$) for a student who scores well, regardless of whether he knows how to add or gets help, it is clear that the posterior distribution*

$$P(x_i) = \frac{L(d|x_i)}{\sum_{j=1}^3 L(d|x_j)} = \frac{\binom{10}{d}\pi_i^d(1-\pi_i)^{10-d}}{\sum_{j=1}^3 \binom{10}{d}\pi_j^d(1-\pi_j)^{10-d}}$$

satisfies $P(x_2) = P(x_3)$. Since the teacher cannot distinguish x_2 from x_3, his sigma-algebra

$$\mathcal{G} = \{\emptyset, \{x_1\}, \{x_2, x_3\}, \mathcal{X}\} \tag{12}$$

has only four elements, whereas the full sigma-algebra $\mathcal{F} = 2^{\mathcal{X}}$ consists of all eight subsets of $\mathcal{X}$. Note that Equation (3) stipulates that the teacher cannot discern between the elements of $\mathcal{X}$, beyond the limits set by $\mathcal{G}$, better than the ignorant person. In order to verify (3), since there is no sigma-algebra between $\mathcal{G}$ and $\mathcal{F}$, we only need to check this equation for $\mathcal{G}' = \mathcal{G}$. To this end, let $g : \mathcal{X} \to \mathbb{R}$ be a real-valued function. Then, since $P(x_2) = P(x_3)$, it follows that

$$E_P(g|\mathcal{G})(x_i) = E_{P_0}(g|\mathcal{G})(x_i) = \begin{cases} g(x_1), & i = 1, \\ (g(x_2) + g(x_3))/2, & i = 2, 3. \end{cases}$$

in agreement with (3).

Illustration 3. *During the Russo-Japanese war, the czar Nicholas II was convinced that Russia would easily defeat Japan [31]. His own biases (he considered the Japanese weak and the Russians superior) and the partial information he received from his advisors blinded him to reality. In the end, Russian forces were heavily beaten by the Japanese. In this scenario, the proposition S is "Russia will beat Japan", $\mathcal{X}$ consists of all possible future scenarios, and $f(x) = 1$ for those scenarios $x \in \mathcal{X}$ in which Russia would win the war. As history reveals, $f(x_0) = 0$. The information he received from his advisors was D, and we know it was heavily biased. Nicholas II adopted (very subjectively!) a correct-proposition-excluding probability measure, as in (11), because he did not even consider the possibility of Russia being defeated. The main reason was a dramatically poor assessment of the likelihood $L(D \mid x)$, for $x \in \mathcal{X}$, on top of a prior P_0 that had a low probability for scenarios $x \in A^c$. Nicholas II's verdict was $P(A) \approx 1$.*

3.3. Expected Posterior Beliefs

Since D is random, so is P. For this reason, the expected posterior distribution

$$\bar{P}(B) = E_{x_0}[Q(B \mid D)] = E_{x_0}[P(B)], \quad \forall B \in \mathcal{F}, \tag{13}$$

will be used occasionally, with an expectation corresponding to averaging over all possible data sets D according to its distribution $L(\cdot|x_0)$ in the true world. Consequently, $\bar{P}(A)$ represents the agent's expected belief in S in the true world x_0. Note in particular that in contrast to the posterior P, the expected posterior $\bar{P}$ is not a purely Bayesian notion, since it depends on x_0.

4. Learning

4.1. Active Information for Quantifying the Amount of Learning

The active information (AIN) of an event B is

$$I^+(B) = \log \frac{P(B)}{P_0(B)}. \tag{14}$$

In particular, $I^+(A)$ quantifies how much an agent has learned about whether S is true or not compared to an ignorant person. By inserting (7) into (14), we find that the AIN

$$I^+(A) = \log \frac{L(D \mid A)}{L(D)} = \log \frac{\int L(D \mid x) dP_0(x \mid A)}{\int L(D \mid x) dP_0(x)} \tag{15}$$

is the logarithm of the ratio between how likely it is to observe data when S holds, and how likely data are when no assumption regarding S is made (see also [32]). The corresponding AIN for expected degrees of beliefs is

$$\bar{I}^+(A) = \log \frac{\bar{P}(A)}{P_0(A)}. \tag{16}$$

Definition 1 (Learning). *Learning about S has occurred (conditionally on observed D) if the probability measure P either satisfies $I^+(A) > 0$ when $x_0 \in A$ or $I^+(A) < 0$ when $x_0 \notin A$. In*

particular, full learning corresponds to $I^+(A) = -\log P_0(A)$ *when* $x_0 \in A$ *and* $I^+(A) = -\infty$ *when* $x_0 \notin A$*. Learning is expected to occur if the probability measure* $\bar{P}$ *is such that* $\bar{I}^+(A) > 0$ *when* $x_0 \in A$ *or* $\bar{I}^+(A) < 0$ *when* $x_0 \notin A$*. In particular, full learning is expected to occur if* $\bar{I}^+(A) = -\log P_0(A)$ *when* $x_0 \in A$ *or* $\bar{I}^+(A) = -\infty$ *when* $x_0 \notin A$*.*

Remark 4. *Two extreme scenarios for the active information, when* $x_0 \in A$*, are*

$$I^+(A) \overset{x_0 \in A}{=} \begin{cases} -\log P_0(A), & \text{if (8) holds and } A \in \mathcal{G}, \\ -\infty, & \text{if (11) holds.} \end{cases} \tag{17}$$

According to Definition 1, the upper part of (17) represents full learning—that is, $P(A) = 1$*; whereas the lower part corresponds to a maximal amount of false belief about S when* $x_0 \in A$*—that is,* $P(A) = 0$*.*

Remark 5. *Suppose S is a proposition that a certain entity or machine functions; then,* $-\log P_0(A)$ *is the functional information associated with the event A of observing such functioning entity [33–35]. In our context, functional information corresponds to the maximal amount of learning about S when the machine works (*$f(x_0) = 1$*).*

4.2. Learning as Hypothesis Testing

It is possible to view the AIN in (15) as a test statistic for choosing between the two statistical hypotheses

$$\begin{array}{ll} H_0: & S \text{ is true } \iff x_0 \in A, \\ H_1: & S \text{ is false } \iff x_0 \notin A, \end{array} \tag{18}$$

with the null distribution H_0 being rejected (conditionally on observed D) when

$$I^+(A) \leq I \tag{19}$$

for some threshold I [36–38]. Typically, this threshold represents a lower bound of what is considered to be a significant amount of learning when S is true. Note in particular that the framework of the hypothesis test, (18) and (19), is frequentist, although we use Bayesian tools (the prior and posterior distributions) to define the test statistic.

In order to introduce performance measures of how much the agent has learnt, let $\Pr_{x_0}$ refer to a probabilities when data $D \sim L(\cdot|x_0)$ are generated according to what one expects in the true world. The type I and II errors of the test (18) and (19) are then defined as

$$\begin{array}{rcll} \alpha(x_0) & = & \Pr_{x_0}[I^+(A) \leq I], & x_0 \in A, \\ \beta(x_0) & = & \Pr_{x_0}[I^+(A) > I], & x_0 \notin A, \end{array} \tag{20}$$

respectively. Both these error probabilities are functions of x_0, and they quantify how much the agent has learnt about the truth (cf. Figure 1 for an illustration).

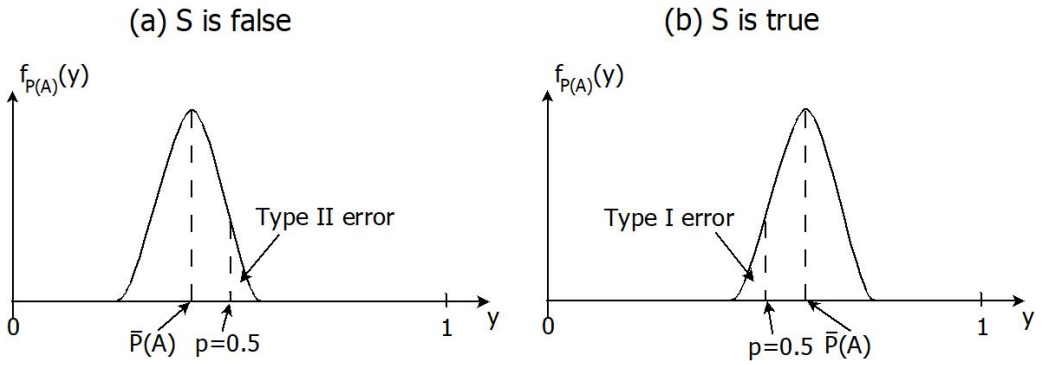

Figure 1. Illustration of the density function $y \to f_{P(A)}(y)$ of $P(A)$ when the data set $D \sim L(\cdot|x_0)$ varies according to the likelihood of the true world parameter for two scenarios where S is either false (**a**) or true (**b**). The threshold of the hypothesis test (19) is $I^+ = \log[p/P_0(A)]$, so that H_0 is rejected when $P(A) \leq p = 0.5$. Note that $\bar{P}(A)$ is the expected value of each density, whereas the error probabilities of type I and II correspond to the areas under the curves in (**b**) and (**a**) to the left and right of p, respectively.

4.3. The Bayesian Approach to Learning

Within a Bayesian framework, we think of H_0 and H_1 as two different models, A and A^c, that represent a subdivision of the parameter space into two disjoints subsets. The posterior odds

$$\text{PostOdds} = \frac{1 - P(A)}{P(A)} = \frac{1 - P_0(A)}{P_0(A)} \cdot \frac{L(D|A^c)}{L(D|A)} = \frac{1 - P_0(A)}{P_0(A)} \cdot \text{BF} \tag{21}$$

factor into a product of the prior odds and the Bayes factor. Hypothesis H_1 is chosen whenever

$$\text{PostOdds} \geq r, \tag{22}$$

for some threshold r. If the cost of drawing a parameter $X \sim P$ from A (A^c) is C_0 (C_1) when H_1 (H_0) is chosen, the optimal Bayesian decision rule corresponds to $r = C_0/C_1$. A little algebra reveals that the AIN is a monotone decreasing function

$$I^+ = -\log[P_0(A)(1 + \text{PostOdds})]$$

of the posterior odds. From this, it follows that the frequentist test (19), with AIN as test statistic, is equivalent to the Bayesian test (22), whenever $I = -\log[P_0(A)(1+r)]$. However, the interpretation of the two tests differ. Whereas the aim of the Bayesian decision rule is to minimize an expected cost (or maximize an expected reward/utility), the aim of the frequentist test is to keep the error probabilities of type I and II low.

4.4. Test Statistic When x_0 Is Unknown

Recall that the agent may or may not know the set A. In the latter case, the agent cannot determine the value of the test statistic $I^+(A)$, and hence he cannot test between H_0 and H_1 himself. This happens, for instance, for the truth function (1), with $A = \{x_0\}$, since the AIN $I^+(A) = \log[p(x_0)/p_0(x_0)]$ then involves the unknown x_0, with $p(x)dx = dP(x)$ and $p_0(x)dx = dP_0(x)$. Although $I^+(A)$ is not known for this particular choice of A, the agent may still use the posterior distribution (7) in order to compute the expected value (conditionally on observed D)

$$\begin{aligned} E_Q[I^+(\{X\})|D] &= E_Q\left[\log \tfrac{p(X)}{p_0(X)}|D\right] \\ &= E_P\left[\log \tfrac{p(X)}{p_0(X)}\right] \\ &= \textstyle\int_{\mathcal{X}} \log \tfrac{p(x)}{p_0(x)} p(x)dx \\ &= D_{KL}(P\|P_0) \\ &= H(P,P_0) - H(P) \end{aligned} \tag{23}$$

of the test statistic according to his posterior beliefs. Note that (23) equals the Kullback–Leibler divergence $D_{KL}(P\|P_0)$ between P and P_0, or the difference between the cross entropy $H(P,P_0)$ between P and P_0, and the entropy $H(P)$ of P. If we also take randomness of the data set D into account, and make use of (7), it follows that the expected AIN, for the same choice of A, equals the mutual information

$$\begin{aligned} E_Q[I^+(\{X\})] &= E_Q\left[E_Q\left[\log \tfrac{p(X)}{p_0(X)}|D\right]\right] \\ &= \textstyle\int \log \tfrac{L(d|x)}{L(d)} dQ(x,d) \\ &= \textstyle\int \log \tfrac{q(x,d)}{p_0(x)L(d)} dQ(x,d), \end{aligned} \tag{24}$$

between X and D, when $(X,D) \sim Q$ vary jointly according to the agent's Bayesian expectations, and with $q(x,d) = dQ(x,d)/d(x,d)$.

5. Knowledge Acquisition

5.1. Knowledge Acquisition Goes beyond Learning

As mentioned in the introduction, knowledge acquisition goes beyond learning, since it also requires that a true belief in S is justified (see Figure 2 for an illustration).

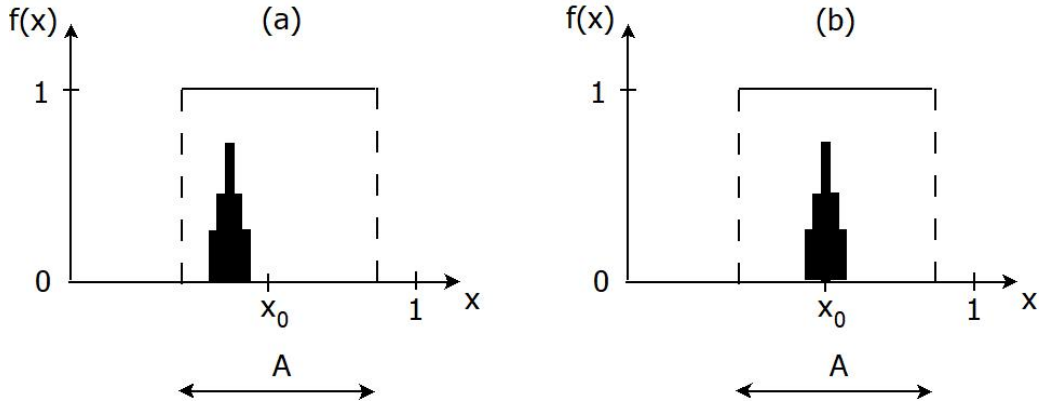

Figure 2. Illustration of the difference between learning and knowledge acquisition for a scenario with a set of worlds $\mathcal{X} = [0,1]$ and a statement S whose truth function $x \to f(x)$ is depicted to the left (**a**) and right (**b**). It is assumed that S is true ($x_0 \in A$), and that the degrees of beliefs P_0 of an ignorant person correspond to a uniform distribution on $\mathcal{X}$. The filled histograms correspond to the density functions $p(x)dx = dP(dx)$ of two agent's beliefs. The agent to the left (**a**) has learnt about S but not acquired knowledge, since x_0 does not belong to the support of P. The agent to the right has not only learnt about S, but also acquired knowledge, since his belief is justified, corresponding to a distribution P that is more concentrated around the true world x_0, compared to the ignorant person. Hence, the JTB condition is satisfied for the agent to the right, but not for the agent to the left.

It is possible, in principle, for an agent whose probability measure P corresponds to a smaller belief in x_0 compared to that of the ignorant person, to have a value of I^+ anywhere in the range $[-\infty, -\log P_0(A)]$ when S is true (i.e., when $x_0 \in A$). One can think of a case in which the agent will believe in S with certainty ($P(A) = 1$) if $\text{supp}(P) \subset A$; but this belief in S is for the wrong reason if, for instance, the agent does not believe in the true world, i.e., if (10) holds, corresponding to the left part of Figure 2. Another less extreme situation

occurs when the agent has a higher belief in A compared to the ignorant person but has lost some (but not all) confidence in the true world with respect to that of the ignorant person; in this case, the agent has not acquired new knowledge about the true world compared to the ignorant person, although he still has learned about S and has some knowledge about the true world.

5.2. A Formal Definition of Knowledge Acquisition

Knowledge acquisition is formulated using tools from statistical estimation theory. Loosely speaking, the agent acquires knowledge, based on data D, if the posterior distribution P gets more concentrated around x_0, compared to an *ignorant* person. By this we mean that each closed ball centered at x_0 has a probability that is at least as large under P as under P_0. Closed balls require, in turn, the concept of a metric or distance; that is, a function $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$. Some examples of metric are:

1. If $\mathcal{X} \subset \mathbb{R}^q$, we use the Euclidean distance $d(x_1, x_2) = \sqrt{\sum_{i=1}^{q} (x_{2i} - x_{1i})^2}$ between $x_1, x_2 \in \mathcal{X}$ as metric.
2. If $\mathcal{X} = \{0, 1\}^q$ consists of all binary sequences of length q, then $d(x_1, x_2) = \sum_{i=1}^{q} |x_{2i} - x_{1i}|$ is the Hamming distance between x_1 and x_2.
3. If $\mathcal{X}$ is a finite categorical space, we put

$$d(x_1, x_2) = \begin{cases} 0, & x_1 = x_2, \\ 1, & x_1 \neq x_2. \end{cases}$$

Equipped with a metric on $\mathcal{X}$, knowledge acquisition is now defined:

Definition 2 (Knowledge acquisition and full knowledge)**.** *Let $B_\epsilon(x_0) = \{x \in \mathcal{X} : d(x, x_0) \leq \epsilon\}$ be the closed ball of radius ϵ that is centered at x_0 with respect to some metric d. We say that an agent has acquired knowledge about S (conditionally on observed D) if learning has occurred according to Definition 1, and in order for this learning to be justified, the following two properties are satisfied for all $\epsilon > 0$:*

$$P(B_\epsilon(x_0)) > 0, \tag{25}$$

and

$$P(B_\epsilon(x_0)) \geq P_0(B_\epsilon(x_0)) \tag{26}$$

with strict inequality for at least one $\epsilon > 0$. Full knowledge about S requires that (9) holds; i.e., that the agent with certainty believes that the true world x_0 is true. The agent is expected to acquire knowledge about S if learning is expected to occur, according to Definition 1, and if (25) and (26) hold with $\bar{P}$ instead of P. The agent is expected to acquire full knowledge about S if (9) holds with $\bar{P}$ instead of P.

Several remarks are in order.

Remark 6. *Property (25) ensures that x_0 is in support of P ([39], p. 20) Kallenberg2021a. When P_0 is the uniform distribution (2), (25) follows from (26). Property (26) is equivalent to $I^+(B_\epsilon(x_0)) \geq 0$, when $P_0(B_\epsilon(x_0)) > 0$. In this case, the requirement that (26) is satisfied with strict inequality for some $\epsilon = \epsilon^* > 0$ is equivalent to learning the proposition S_{ϵ^*}: "The distance of a world to the true world x_0 is less than or equal to ϵ^*," corresponding to a truth function*

$$f_{\epsilon^*}(x) = 1(x \in B_{\epsilon^*}(x_0)). \tag{27}$$

Since the agent does not know x_0, neither f_{ϵ^*} nor $A_{\epsilon^*} = \{x \in \mathcal{X}; f_{\epsilon^*}(x) = 1\}$ is known to him, even if he is able to discern between all possible worlds. If f_{ϵ^*} differs from the original truth function f, learning of S_{ϵ^*} can be viewed as meta-learning. Note also that $A_0 = \{x_0\}$ corresponds to the set in (1).

Remark 7. *Suppose the truth function used to define learning and knowledge acquisition satisfies (27), i.e., $f = f_\epsilon$ for some $\epsilon \geq 0$. Then (25) and (26) are sufficient for knowledge acquisition, since they imply that learning of $S = S_{\epsilon^*}$ has occurred, according to Definition 1. Although knowledge acquisition in general requires more than learning, the two concepts are equivalent for a truth function $f = f_0$, with $A = A_0 = \{x_0\}$, as defined in (1). Indeed, in this case it is not possible to learn whether $S = S_0$ is true or not for the wrong reason.*

Remark 8. *Recall from Definition 1 that an agent has fully learnt S when*

$$P(A) = \mathbf{1}(x_0 \in A) = \begin{cases} 1, & x_0 \in A, \\ 0, & x_0 \notin A. \end{cases} \tag{28}$$

For a rational agent, the lower part of (28) should hold when data D falsifies S. In general, (28) is a necessary but not sufficient condition for full knowledge. Indeed, it follows from (9) that, for a person to have full knowledge, $P(B) = \mathbf{1}(x_0 \in B)$ must hold for all $B \in \mathcal{F}$, not only for the set A of worlds for which S is true.

Remark 9. *Suppose a distance measure $d(P, Q)$ between probability distributions on $(\mathcal{X}, \mathcal{F})$ is defined. This gives rise to a different definition of knowledge acquisition, whereby the agent acquires knowledge if has learnt about S and additionally $d(P, \delta_{x_0}) < d(P_0, \delta_{x_0})$, that is, if his beliefs are closer than the ignorant person's beliefs to the Oracle's beliefs. Possible choices of distances are the Kullback–Leibler divergence $d(P, Q) = D_{KL}(Q||P)$ and the Wasserstein metric $d(P, Q) = \min_{X_1, X_2} E|X_1 - X_2|$, where the minimum is taken over all random vectors (X_1, X_2) whose marginals have distributions P and Q, respectively. Note in particular that the KL choice of distance yields $d(Q, \delta_{x_0}) = -\log Q(x_0)$. The corresponding notion of knowledge acquisition is weaker than in Definition 2, requiring (25) and (26) to hold only for $\epsilon = 0$.*

Illustration 4 (Continuation of Illustration 1)**.** *To check whether learning or knowledge acquisition has occurred, according to Definitions 1 and 2, for the student who takes the math home assignment, x_0 must be known. The reader may think of an instructor with full information—an $\mathcal{F}$-optimal measure according to (9)—who checks whether a pupil has learned and acquired knowledge or not. However, in Illustration 1 it is the teacher who is the pupil and learns and acquires knowledge about the skills of a math student. In this context, the instructor is a supervisor of the teacher who knows whether the math student is able to add ($x_0 = x_2$) or not, and in the latter case whether the student gets help ($x_0 = x_3$) or not ($x_0 = x_1$). Whereas the instructor's sigma-algebra is $\mathcal{F}$, the teacher's sigma-algebra $\mathcal{G}$ in (12) does not make it possible to discern between x_2 and x_3. Suppose $x_0 = x_2$. No matter how many questions the home exam has, as long as the teacher does not get information from the instructor on whether the student solved the home exam without help or not, although the teacher learns that S is true, since the student scores well, he will never acquire full knowledge that the student knows how to add, since $P(x_0) = P(x_2) = P(x_3) \leq 0.5 < 1$.*

6. Learning and Knowledge Acquisition Processes

The previous two sections dealt with learning and knowledge acquisition of a static belief P, corresponding to an agent who is able to discern between worlds according to one sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$, and who has access to one data set D. The setting is now extended to consider an agent who is exposed to an increasing amount of information about (or discernment between) the possible worlds in $\mathcal{X}$, and increasingly larger data sets.

6.1. The Process of Discernment and Data Collection

Mathematically, an increased ability to discern between possible worlds is expressed as a sequence of σ-algebras

$$\mathcal{G}_1 \subset \ldots \subset \mathcal{G}_n \subset \mathcal{F}. \tag{29}$$

Typically, $\mathcal{G}_k$ is generated by a random variable Y_k whose domain is in $\mathcal{X}$ for $k = 1, \ldots, n$. The σ-algebras in (29) are associated with increasingly larger data sets $D_1, \ldots, D_n$, with

$D_k \in \mathcal{D}_k$. Let $dQ_k(x, D_k) = dP_0(x)L(D_k|x)dD_k$ refer to the joint distribution of the parameter and data in step k, such that the likelihood $x \to L(D_k \mid x)$ of D_k is $\mathcal{G}_k$-measurable. This implies that an agent who interprets data D_k according to this likelihood function has beliefs (represented by the posterior probability measure $P_k(\cdot) = Q_k(\,\cdot \mid D_k)$) that correspond to not being able to discern events outside of $\mathcal{G}_k$ better than an ignorant person. Mathematically, this is phrased as a requirement

$$E_{P_k}[g \mid \mathcal{G}'_k] = E_{P_0}[g \mid \mathcal{G}'_k], \tag{30}$$

for all $\mathcal{F}$-measurable functions $g : \mathcal{X} \to \mathbb{R}$ and sigma algebras $\mathcal{G}'_k$ such that $\mathcal{G}_k \subset \mathcal{G}'_k \subset \mathcal{F}$, for $k = 1, \ldots, n$. The collection of pairs $(D_1, \mathcal{G}_1), \ldots, (D_n, \mathcal{G}_n)$ is referred to as a discernment and data collection process. The active information, after k steps of the discernment and data collection process, is

$$I_k^+(A) = \log \frac{P_k(A)}{P_0(A)}. \tag{31}$$

Let $\bar{P}_k(\cdot) = E_{x_0}[P_k(\,\cdot \mid D_k)]$ refer to expected degrees of belief after k steps of the information and data collection process, if data $D_k \sim L(\cdot|x_0)$ vary according that what one expects in the true world. The corresponding active information is

$$\bar{I}_k^+(A) = \log \frac{\bar{P}_k(A)}{P_0(A)}. \tag{32}$$

In the following sections we will use the sequences $I_1^+, \ldots, I_n^+$ and $P_1, \ldots, P_n$ of AINs and posterior beliefs in order to define different notions of learning and knowledge acquisition.

6.2. Strong Learning and Knowledge Acquisition

Definition 3 (Strong learning)**.** *The probability measures $P_1, \ldots, P_n$, obtained from the discernment and data collection process represent a learning process in the strong sense (conditionally on observed $D_1, \ldots, D_n$) if*

$$\begin{cases} 0 \le I_1^+(A) \le \cdots \le I_n^+(A), & \text{if } x_0 \in A, \\ 0 \ge I_1^+(A) \ge \cdots \ge I_n^+(A), & \text{if } x_0 \notin A, \end{cases} \tag{33}$$

with at least one strict inequality. Learning is expected to occur, in the strong sense, if (33) holds with $\bar{I}_1^+(A), \ldots, \bar{I}_n^+(A)$, instead of $I_1^+(A), \ldots, I_n^+(A)$.

Definition 4 (Strong knowledge acquisition)**.** *With $B_\epsilon(x_0)$ as in Definition 2, the learning process is knowledge acquiring in the strong sense (conditionally on observed $D_1, \ldots, D_n$) if, in addition to (33), we have that this learning process is justified, so that for all $\epsilon > 0$, $P_1(B_\epsilon(x_0)) > 0$ and*

$$P_0(B_\epsilon(x_0)) \le P_1(B_\epsilon(x_0)) \le \cdots \le P_n(B_\epsilon(x_0)), \tag{34}$$

with strict inequality for at least one step of (34) and for at least one $\epsilon > 0$. Knowledge acquisition is expected to occur, in the strong sense, if learning is expected to occur in the strong sense, according to Definition 3, and additionally (34) holds with $\bar{P}_1, \ldots, \bar{P}_n$, instead of $P_1, \ldots, P_n$.

Illustration 5 (Continuation of Illustration 1)**.** *Assume the teacher of the math student has a discernment and data collection process $(\mathcal{G}_1, D_1), (\mathcal{G}_2, D_2)$, where in the first step, $\mathcal{G}_1 = \mathcal{G}$ and $D_1|x_i \sim Bin(10, \pi_i)$ are obtained from a home assignment with 10 questions (as described in Section 3.2). Suppose the student knows how to add ($x_0 = x_2$). It can be seen that*

$$\begin{array}{rcl} P_1(A) & = & P_1(x_2) + P_1(x_3) > P_0(A) = 2/3, \\ P_1(x_0) & = & P_1(x_2) > P_0(x_2) = 1/3 \end{array} \tag{35}$$

whenever $7 \leq d_1 \leq 10$. *Assume that in a second step the teacher receives information* $Z_2 \in \{0,1\}$ *from the instructor on whether the student used external help* ($Z_2 = 1$) *or not* ($Z_2 = 0$) *during the exam. Let* $d_2 = (d_1, z_2)$ *refer to observed data after step 2. The likelihood, after the second step, then takes the form*

$$L(d_2|x_i) = \binom{10}{d_1} \pi_i^{d_1} (1-\pi_i)^{10-d_i} \cdot L(z_2|x_i),$$

where $L(1|x_i) = 1(x_i = x_3)$ *and* $L(0|x_i) = 1(x_i \in \{x_1, x_2\})$. *If the instructor correctly reports that the student did not use external help* ($z_2 = 0$), *it follows that*

$$\begin{array}{rcl} P_2(A) & = & P_2(x_2) = P_1(x_2)/(P_1(x_1) + P_1(x_2)) < 2P_1(x_2)/(P_1(x_1) + 2P_1(x_2)) = P_1(A), \\ P_2(x_0) & = & P_2(x_2) > P_1(x_2)/(P_1(x_1) + 2P_1(x_2)) = P_1(x_2) = P_1(x_0). \end{array} \tag{36}$$

We deduce from (35) *and* (36) *that*

$$P_0(x_0) < P_1(x_0) < P_2(x_0), \tag{37}$$

which suggests that knowledge acquisition has occurred if the categorical space metric $d(x_i, x_j) = 1(x_i \neq x_j)$ *is used on* $\mathcal{X}$. *However, since* $P_2(A) < P_1(A)$, *neither learning nor knowledge acquisition in the strong sense has occurred. The reason is that the information from the instructor (that the student has not cheated) makes the teacher less certain as to whether the student is able to score well on the test. On the hand, if we change the proposition to S: "The student knows how to add," with* $A = \{x_2\}$, *then strong learning and knowledge acquisition has occurred because of* (37), *since* $P_k(A) = P_k(x_0)$ *for* $k = 0, 1, 2$.

6.3. Weak Learning and Knowledge Acquisition

Learning and knowledge acquisition are often fluctuating processes, and the requirements of Definition 3 are sometimes too strict. Accordingly, weaker versions of learning and knowledge acquisition are thus introduced.

Definition 5 (Weak learning). *Learning in the weak sense has occurred at time n (conditionally on observed* D_n), *if*

$$\begin{cases} 0 < I_n^+(A), & \text{if } x_0 \in A, \\ 0 > I_n^+(A), & \text{if } x_0 \notin A. \end{cases} \tag{38}$$

Learning is expected to occur in the weak sense if (38) *holds with* $\bar{I}_n^+$ *instead of* I_n^+.

Definition 6 (Weak knowledge acquisition). *Knowledge acquisition in the weak sense occurs (conditionally on observed* D_n) *if, in addition to the weak learning condition* (38), *in order for this learning to be justified, it holds for all* $\epsilon > 0$ *that* $P_n(B_\epsilon(x_0)) > 0$ *and*

$$P_0(B_\epsilon(x_0)) \leq P_n(B_\epsilon(x_0)), \tag{39}$$

with strict inequality for at least one $\epsilon > 0$. *Knowledge acquisition is expected to occur in the weak sense if weak learning occurs according to Definition 5 and* (39) *holds with* $\bar{P}_n$ *instead of* P_n.

7. Asymptotics

Strong and weak learning (or strong and weak knowledge acquisition) are equivalent for $n = 1$. The larger n is, the more restrictive strong learning becomes in comparison to weak learning. However, for large n, neither strong nor weak learning (knowledge acquisition) are entirely satisfactory entities. For this reason, in this section we will introduce asymptotic versions of learning and knowledge acquisition, for an agent whose discernment between worlds and collected data sets increase over a long period of time.

7.1. Asymptotic Learning and Knowledge Acquisition

In order to define asymptotic learning and knowledge acquisition, as the number of steps n of the discernment and data collection process tends to infinity, we first need to introduce AIN versions of limits. Define

$$I^+_{\liminf}(B) = \log \frac{\liminf P_k(B)}{P_0(B)}, \tag{40}$$

$$I^+_{\limsup}(B) = \log \frac{\limsup P_k(B)}{P_0(B)}, \tag{41}$$

and when the two limits of (40) agree, we refer to the common value as $I^+_{\lim}(B)$. Define also

$$\bar{I}^+_{\liminf}(B) = \log \frac{\liminf \bar{P}_k(B)}{P_0(B)}, \tag{42}$$

$$\bar{I}^+_{\limsup}(B) = \log \frac{\limsup \bar{P}_k(B)}{P_0(B)}, \tag{43}$$

with $\bar{I}^+_{\lim}(B)$ the common value whenever the two limits of (42) agree. Since $I^+_{\lim}(B)$ only exists when $I^+_{\liminf}(B) = I^+_{\limsup}(B)$, and $I^+_{\liminf}(B) \leq I^+_{\limsup}(B)$, the following definitions of asymptotic learning and knowledge acquisition are natural:

Definition 7 (Asymptotic learning). *Learning occurs asymptotically (conditionally on the observed data sequence $\{D_k\}_{k=1}^{\infty}$) if*

$$\begin{cases} I^+_{\liminf}(A) > 0, \text{ for } x_0 \in A, \\ I^+_{\limsup}(A) < 0, \text{ for } x_0 \notin A. \end{cases} \tag{44}$$

Full learning occurs asymptotically (conditionally on $\{D_k\}_{k=1}^{\infty}$) if

$$\begin{cases} I^+_{\lim}(A) = -\log P_0(A), \text{ for } x_0 \in A, \\ I^+_{\lim}(A) = -\infty, \text{ for } x_0 \notin A. \end{cases} \tag{45}$$

Learning is expected to occur asymptotically if (44) *holds with $\bar{I}^+_{\limsup}$ and $\bar{I}^+_{\liminf}$, instead of $I^+_{\limsup}$ and $I^+_{\liminf}$, respectively. Full learning is expected to occur asymptotically, if* (45) *holds with $\bar{I}^+_{\lim}$ instead of $I^+_{\lim}$.*

Definition 8 (Asymptotic knowledge acquisition). *Knowledge acquisition occurs asymptotically (conditionally on $\{D_k\}_{k=1}^{\infty}$) if, in addition to the asymptotic learning condition* (44)*, in order for this asymptotic learning to be justified, for every $\epsilon > 0$, it holds that*

$$\liminf_{k\to\infty} P_k(B_\epsilon(x_0)) > 0$$

and

$$I^+_{\liminf}(B_\epsilon(x_0)) \geq 0, \tag{46}$$

with strict inequality for a least one $\epsilon > 0$. Full knowledge acquisition occurs asymptotically (conditionally on $\{D_k\}_{k=1}^{\infty}$) if (45) *holds and*

$$I^+_{\lim}(B_\epsilon(x_0)) = -\log P_0(B_\epsilon(x_0)) \tag{47}$$

is satisfied for all $\epsilon > 0$. If learning is expected to occur asymptotically according to Definition 7, and if (46) *holds with $\bar{I}^+_{\liminf}$ instead of $I^+_{\liminf}$, then knowledge acquisition is expected to occur asymptotically. Full knowledge acquisition is expected to occur asymptotically if full learning is expected to occur asymptotically according to Definition 7, and if* (47) *holds with $\bar{I}^+_{\lim}$ instead of $I^+_{\lim}$.*

7.2. Bayesian Asymptotic Theory

In this subsection we will use Bayesian asymptotic theory in order to quantify and give conditions for when asymptotic learning and knowledge acquisition occur. Let Ω be a large space that incorporates prior beliefs and data for all $k = 1, 2, \ldots$. Define $X_k : \Omega \to \mathcal{X}$ as a random variable whose distribution corresponds to the agent's posterior beliefs, based on data set D_k, which itself varies according to another random variable $D_k : \Omega \to \mathcal{D}_k$ with distribution $D_k \sim L(\cdot|x_0)$. Let $\Pr_{x_0}$ be a probability measure on subsets of Ω that induces distributions $X_k|D_k \sim P_k$ and $X_k \sim \bar{P}_k$, respectively. The following proposition is a consequence of Definitions 7 and 8:

Proposition 1. *Suppose full learning is expected to occur asymptotically, in the sense of* (45)*, with* $\bar{I}^+_{\lim}$ *instead of* $I^+_{\lim}$*. Then,*

$$Pr_{x_0}(X_k \in A) = \bar{P}_k(A) \to \begin{cases} 1, & x_0 \in A, \\ 0, & x_0 \notin A \end{cases} \tag{48}$$

as $k \to \infty$*. In particular, the type I and II errors of the hypothesis test* (18) *and* (19)*, with threshold* $I = \log[p/P_0(A)]$ *for some* $0 < p < 1$*, satisfy*

$$\begin{array}{rcl} \alpha_k(x_0) & = & Pr_{x_0}\left(I^+_k(A) \leq I\right) = Pr_{x_0}[Pr(X_k \in A \mid D_k) \leq p] \\ & = & Pr_{x_0}(P_k(A) \leq p) \to 0, \quad x_0 \in A, \\ \beta_k(x_0) & = & Pr_{x_0}\left(I^+_k(A) > I\right) = Pr_{x_0}[Pr(X_k \in A \mid D_k) > p] \\ & = & Pr_{x_0}(P_k(A) > p) \to 0, \quad x_0 \notin A, \end{array} \tag{49}$$

respectively, as $k \to \infty$*. If full knowledge acquisition occurs asymptotically, in the sense of* (47)*, then*

$$X_k \mid D_k \xrightarrow{p} x_0 \text{ conditionally on } \{D_k\}_{k=1}^{\infty} \tag{50}$$

as $k \to \infty$*, with* $\xrightarrow{p}$ *referring to convergence in probability. If full knowledge acquisition is expected to occur asymptotically, in the sense of Definition 8, then*

$$X_k \xrightarrow{p} x_0 \tag{51}$$

as $k \to \infty$*.*

Remark 10. *Full asymptotic knowledge acquisition* (50) *is closely related to the notion of posterior consistency [40]. For our model, the latter concept is usually defined as*

$$Pr_{x_0}\left(X_k \mid D_k \xrightarrow{p} x_0 \text{ as } k \to \infty\right) = 1, \tag{52}$$

where the probability refers to variations in the data sequence $\{D_k\}_{k=1}^{\infty}$ *when* x_0 *holds. Thus, posterior consistency* (52) *means that full asymptotic knowledge acquisition* (50) *holds with probability 1. Let* $\mathcal{L}(X)$ *refer to the distribution of the random variable* X*. Then,* (52) *is equivalent to*

$$P_k = \mathcal{L}(X_k \mid D_k) \xrightarrow{\text{a.s.}} \delta_{x_0} \tag{53}$$

as $k \to \infty$*, with* $\xrightarrow{\text{a.s.}}$ *referring to almost sure weak convergence with respect to variations in the data sequence* $\{D_k\}_{k=1}^{\infty}$ *when* $x = x_0$*. On the other hand, it follows from Definition 8 that if full knowledge acquisition is expected to occur asymptotically, this is equivalent to*

$$P_k = \mathcal{L}(X_k \mid D_k) \xrightarrow{p} \delta_{x_0} \tag{54}$$

as $k \to \infty$*, which is a weaker concept than posterior consistency, since almost sure weak convergence implies weak convergence in probability. However, sometimes* (54)*, rather than* (52) *and* (53)*, is used as a definition of posterior consistency.*

Remark 11. *It is sometimes possible to sharpen* (54) *and obtain the rate at which the posterior distribution converges to* δ_{x_0}. *The posterior distribution is said to contract at rate* $\epsilon_k \to 0$ *to* δ_{x_0} *as* $k \to \infty$ *(see for instance [41]), if for every sequence* $M_k \to \infty$ *it holds that*

$$Q[X_k \mid D_k \notin B(x_0, M_k\epsilon_k)] = P_k[B(x_0, M_k\epsilon_k)^c] \xrightarrow{p} 0, \tag{55}$$

when $\{D_k\}_{k=1}^{\infty}$ *varies according to what one expects in the true world* x_0. *Since convergence in probability is equivalent to convergence in mean for bounded random variables, it can be seen that* (54) *is equivalent to* $\bar{P}_k[B(x_0, M_k\epsilon_k)^c] \to 0$, *or*

$$(M_k\epsilon_k)^{-1}(X_k - x_0) \xrightarrow{p} 0 \tag{56}$$

as $k \to \infty$ *for each sequence* $M_k \to \infty$. *Comparing* (51) *with* (55) *and* (56), *we found that a contraction of the posterior towards* δ_{x_0} *at rate* ϵ_k *is equivalent to expecting full knowledge acquisition asymptotically at rate* ϵ_k.

It follows from Proposition 1 and Remarks 10 and 11 that Bayesian asymptotic theory can be used, within our frequentist/Bayesian framework, to give sufficient conditions for asymptotic learning and knowledge acquisition to occur. Suppose, for instance, that $D_k = (Z_1, \ldots, Z_k)$ is a sample of k independent and identically distributed random variables Z_l with distribution $Z_l \sim F(\cdot \mid x_0)$ that belongs to the statistical model $\{F(\cdot \mid x);\, x \in \mathcal{X}\}$. The likelihood function is then a product

$$L(D_k \mid x) = \prod_{l=1}^{k} \Pr(Z_l \mid x) \tag{57}$$

of the likelihoods of all observations Z_l. For such a model, a number of authors [40,42–45] have provided sufficient conditions for posterior consistency (52) and (53) to occur. It follows from Remark 10 that these conditions also imply the weaker concept (54) of full, expected knowledge acquisition to occur asymptotically.

Suppose (57) holds with a parameter space $\mathcal{X} \subset \mathbb{R}^q$ that is a subset of Euclidean space of dimension q. It is possible then to obtain the rate (56) at which knowledge acquisition is expected to occur. The first step is to use the Bernstein–von Mises theorem, which under appropriate conditions (see for instance [46]) approximates the posterior distribution $P_k = \mathcal{L}(X_k \mid D_k)$ by a normal distribution centered around the maximum likelihood (ML) estimator

$$\hat{x}_{0k} = \hat{x}_{0k}(D_k) = \operatorname{argmax}_{x \in \mathcal{X}} L(D_k \mid x) \tag{58}$$

of x_0. More specifically, this theorem provides weak convergence

$$\sqrt{k}(X_k - \hat{x}_{0k}) | D_k \xrightarrow{\mathcal{L}} N\left(0, J(x_0)^{-1}\right) \tag{59}$$

as $k \to \infty$, of a re-scaled version of the distribution of $X_k | D_k$ when $\{D_k\}_{k=1}^{\infty}$ varies according what one expects when $x = x_0$. The limiting distribution is a q-dimensional normal distribution with mean 0 and a covariance matrix that equals the inverse of the Fisher information matrix $J(x_0)$, evaluated at the true world parameter x_0. On the other hand, the standard asymptotic theory of maximum likelihood estimation (see for instance [47]) implies

$$\sqrt{k}(\hat{x}_{0k} - x_0) \xrightarrow{\mathcal{L}} N\left(0, J(x_0)^{-1}\right) \tag{60}$$

as $k \to \infty$, with weak convergence referring to variations in the data sequence $\{D_k\}_{k=1}^{\infty}$ when $x = x_0$. Combining equations (59) and (60), we arrive at the following result:

Theorem 1. *Assume data* $\{D_k = (Z_1, \ldots, Z_k)\}_{k=1}^{\infty}$ *consists of independent and identically distributed random variables* $\{Z_l\}_{l=1}^{\infty}$, *and that the Bernstein–von Mises theorem* (59) *and asymptotic*

normality (60) *of the ML-estimator hold. Then,* X_k *converges weakly towards* x_0 *at rate* $1/\sqrt{k}$*, in the sense that*

$$\sqrt{k}(X_k - x_0) \xrightarrow{\mathcal{L}} N\Big(0, 2J(x_0)^{-1}\Big) \tag{61}$$

as $k \to \infty$*. In particular, full knowledge acquisition is expected to occur asymptotically at rate* $1/\sqrt{k}$*.*

Proof. Let $s \to F_k(s) = F_{\sqrt{k}(\hat{x}_{k0} - x_0)}(s)$ refer to the distribution function of $\sqrt{k}(\hat{x}_{k0}(D_k) - x_0)$, defined for all q-dimensional vectors $s = (s_1, \ldots, s_q) \in \mathbb{R}^q$. Let also $t = (t_1, \ldots, t_q) \in R^q$ and denote the distribution function of $N\big(0, 2J(x_0)^{-1}\big)$ by G. Combining (59) and (60), and making use of the fact that the convolution of two independent $N\big(0, J(x_0)^{-1}\big)$-variables is distributed as $N\big(0, 2J(x_0)^{-1}\big)$, we can find that

$$\begin{array}{rcl} \Pr_{x_0}\Big(\sqrt{k}(X_k - x_0) \le t\Big) & = & \int \Pr_{x_0}\Big(\sqrt{k}(X_k - \hat{x}_{0k}) \le t - s | \sqrt{k}(\hat{x}_{0k} - x_0) = s\Big) dF_k(s) \\ & \to & \int_{s \le t} dG(s) \end{array} \tag{62}$$

as $k \to \infty$, with $s \le t$ referring to $s_j \le t_j$ for $j = 1, \ldots, q$. Since (62) holds for any $t \in \mathbb{R}^q$, Equation (61) follows. Moreover, in view of (56), Equation (61) implies that full knowledge acquisition is expected to occur asymptotically at rate $1/\sqrt{k}$. □

In general, the conditions of Theorem 1 typically require that data, and the agent's interpretation of data, are unbiased. When these conditions fail (cf. Remark 2), there is no guarantee that knowledge acquisition is expected to occur asymptotically as $k \to \infty$.

8. Examples

Example 1 (Coin tossing)**.** *Let* $x_0 \in \mathcal{X} = [0, 1]$ *be the probability of heads when a certain coin is tossed. An agent wants to find out whether the proposition*

$$S: \text{ the coin is symmetric with margin } \varepsilon > 0$$

is true or not. This corresponds to a truth function $f(x) = \mathbf{1}(x \in A)$*, with* $A = [0.5 - \varepsilon, 0.5 + \varepsilon]$*, that is known to the agent. Suppose the coin is tossed a large number of times* ($n = \infty$)*, and let* $D_k = (Z_1, \ldots, Z_k) \in \mathcal{D}_k = \{0, 1\}^k$ *be a binary sequence of length* k *that represents the first* k *tosses, with tails and heads corresponding to 0* ($Z_k = 0$) *and 1* ($Z_k = 1$)*, respectively. The number of heads* $M_k = \sum_{l=1}^{k} Z_l \sim Bin(k, x_0)$ *after* k *tosses is then a sufficient statistic for estimating* x_0 *based on data* D_k*. Even though* $\{D_k\}$ *is an increasing sequence of data sets, we put* $Y_k(x) = x$ *and* $\mathcal{G}_k = \mathcal{F} = \mathcal{B}([0, 1])$*, the Borel* σ*-algebra on* $[0, 1]$*, for* $k = 1, 2, \ldots$*. Let* P_0 *be the uniform prior distribution on* $[0, 1]$*. Since the uniform distribution is a beta distribution, and beta distributions are conjugate priors to binomial distributions, it is well known [17] that the posterior distribution*

$$P_k \sim Beta(1 + M_k, 1 + k - M_k)$$

belongs to the beta family as well. Consequently, if X_k *is a random variable that reflects the agent's degree of belief in the probability of heads after* k *tosses, it follows that his belief in a symmetric coin, if* $M_k = m$*, is*

$$\begin{aligned} P_k(A) = Pr(X_k \in A \mid D_k) &= Pr(X_k \in A \mid M_k = m) \\ &= \int_{0.5-\varepsilon}^{0.5+\varepsilon} p_k(x \mid m) dx, \end{aligned}$$

where

$$\begin{aligned} p_k(x \mid m) &= \frac{(1-x)^{k-m}x^m}{B(1+m,1+k-m)} \\ &= (k+1)\binom{k}{m}(1-x)^{k-m}x^m \\ &= (k+1)L(m \mid x) \end{aligned} \tag{63}$$

is the posterior density function of the parameter x, whereas $B(a,b)$ is the beta function and $x \to L(m \mid x)$ the likelihood function. From this, it follows that the AIN after k coin tosses with m heads and $k-m$ tails equals

$$\begin{aligned} I_k^+(A) &= \log[(2\varepsilon)^{-1}P_k(A)] \\ &= \log\Big[(2\varepsilon)^{-1}(k+1)\tbinom{k}{m}\textstyle\int_{0.5-\varepsilon}^{0.5+\varepsilon}(1-x)^{k-m}x^m dx\Big]. \end{aligned}$$

Since data are random, $P_k(A)$ (and hence also $I_k^+(A)$) will fluctuate randomly up and down with probability one (see Figure 3); for this reason, $\{P_k\}_{k=1}^{\infty}$ does not represent a learning process in the strong sense of Definition 3. On the other hand, it follows by the strong law of large numbers that $M_k/k \xrightarrow{a.s.} x_0$ as $k \to \infty$, and from properties of the beta distribution, this implies that full learning and knowledge acquisition occur asymptotically according to Definitions 7 and 8, with probability 1. In view of Remark 10, we also have posterior consistency (52) and (53).

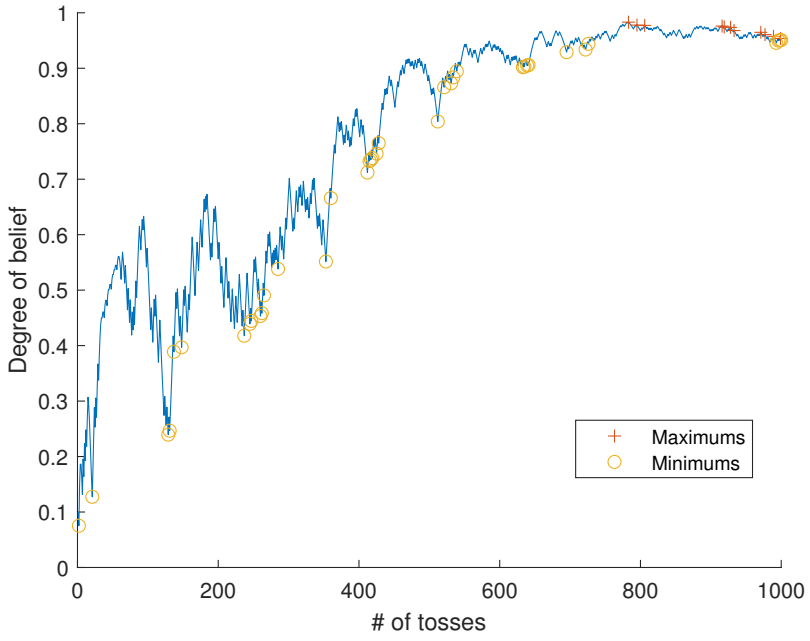

Figure 3. Degree of belief is represented as a function of coin tosses. There is no strong learning because the belief oscillates. However, there is weak learning after a few coin tosses. In particular, when the number of coin tosses is 1000, there is weak learning since $P_{1000}(A) > P_0(A)$ and $I_{1000}^+(A) > 0$.

By analyzing $\bar{P}_k$ instead of P_k, we may also assess whether learning and knowledge acquisition are expected to occur. The expected degree of belief in a symmetric coin, after k tosses, is

$$\bar{P}_k(A) = \int_{0.5-\varepsilon}^{0.5+\varepsilon} \bar{p}_k(x)dx,$$

where

$$\begin{aligned}\bar{p}_k(x) &= E_{x_0}[p_k(x \mid M_k)] \\ &= \sum_{m=0}^{k} L(m \mid x_0) p_k(x \mid m) \\ &= (k+1) \sum_{m=0}^{k} L(m \mid x_0) L(m \mid x)\end{aligned}$$

is the expected posterior density function of x, after k tosses of the coin. Note in particular that

$$\int_0^1 \bar{p}_k(x) dx = 1.$$

It can be shown that (63) and the weak law of large numbers ($M_k/k \xrightarrow{p} x_0$ as $k \to \infty$, where $\xrightarrow{p}$ refers to convergence in probability) lead to uniform convergence

$$\sup_{x:|x-x_0|\geq\epsilon} \bar{p}_k(x) \to 0$$

as $k \to \infty$ for any $\epsilon > 0$. The last four displayed equations imply $\bar{P}_k(A) \to \mathbf{1}(x_0 \in A)$ and $\bar{P}_k \xrightarrow{p} x_0$ as $k \to \infty$. This and Definitions 7 and 8 imply that full learning and knowledge acquisition are expected to occur asymptotically. This result is also a consequence of posterior consistency, or of Theorem 1. Notice, however, that a purely Bayesian analysis does not allow us to conclude that knowledge acquisition occurs, or is expected to occur, asymptotically.

Example 2 (Historical events)**.** *Let $\mathcal{X} = (0,1]$ represent a time interval of the past. A person wants to find out whether his ancestor died or not during a famine that occurred in the province where the ancestor lived. Formally, this corresponds to a proposition*

$$S: \text{ the ancestor died during the time of the famine.}$$

Let $f(x) = 1$ if the famine occurred at time x, and $f(x) = 0$ if not. Assume that the ancestor died at an unknown time point x_0 and that the time period during which the famine lasted is $A = [a,b]$, where $0 \leq a < b \leq 1$ are known. If $\mathcal{X}$ corresponds to a fairly short time interval of the past, it is reasonable to assume that P_0 has a uniform distribution on $(0,1]$.

In the first step of the learning process, suppose radiometric dating $D_1 = Z_1$ of a burial find from the ancestor appears. If $\delta = 1/N$ represents the precision of this dating method, the corresponding σ-algebra is

$$\begin{aligned}\mathcal{G}_1 &= \sigma((0,1/\delta], (1/\delta, 2/\delta], \ldots, [(N-1)/\delta, 1]) \\ &= \sigma(Y_1),\end{aligned} \tag{64}$$

where $Y_1 : \mathcal{X} \to \{1, \ldots, N\}$ is defined through $Y_1(x) = [x/\delta] + 1$, and where $[x/\delta]$ is the integer part of x/δ. Due to (3), it follows that P_1 has a density function

$$p_1(x) = N \sum_{i=1}^{N} p_{1i} \mathbf{1}(x \in ((i-1)\delta, i\delta]), \tag{65}$$

for some non-negative probability weights $p_{1i} \geq 0$ that sum to 1. Since $p_{1i} = p_{1i}(D_1)$, this measure is constructed from the radiometric dating data D_1 of the burial find from the ancestor. The $\mathcal{G}_1$-optimal probability measure is obtained from (8) as

$$p_{1i} = \begin{cases} 1, & i = i_0 = [x_0/\delta] + 1, \\ 0, & i \neq i_0 = [x_0/\delta] + 1. \end{cases}$$

It corresponds to dating the time of death of the ancestor correctly, given the accuracy of this dating method. On the other hand, if the radiometric dating equipment has a systematic error of $-\delta$, a truth-excluding probability measure (10) is obtained with

$$p_{1i} = \begin{cases} 1, & i = i_0 - 1 = [x_0/\delta], \\ 0, & i \neq i_0 - 1 = [x_0/\delta]. \end{cases} \tag{66}$$

In the second step of the learning process, suppose data $D_2 = (Z_1, Z_2)$ is extended to include a piece of text Z_2 from a book where the time of death of the ancestor can be found. This extra source of information increases the σ-algebra to $\mathcal{G}_2 = \mathcal{F} = \mathcal{B}([0,1])$, and if the contents of the book are reliable, $P_2 = \delta_{x_0}$ is $\mathcal{F}$-optimal. It follows from Definition 3 that strong learning has occurred if $Na = i_a$ and $Nb = i_b$ are integers and

$$\begin{aligned} 0 < I_1^+ = \log[\textstyle\sum_{i=i_a+1}^{i_b} p_{1i}/(b-a)] < I_2^+ = \log[1/(b-a)], & \quad \text{if } x_0 \in (a,b), \\ 0 > I_1^+ = \log[\textstyle\sum_{i=i_a+1}^{i_b} p_{1i}/(b-a)] > I_2^+ = -\infty, & \quad \text{if } x_0 \notin (a,b). \end{aligned} \tag{67}$$

Figure 4 illustrates another scenario where not only strong learning but also strong knowledge acquisition occurs. Suppose now that (66) holds, with $i_a + 1 \leq i_0 - 1 \leq i_b$. If $P_2 = \delta_{x_0}$, the strong learning condition (67) is satisfied, and the weak knowledge acquisition requirement of Definition 6 holds as well. Strong knowledge acquisition has not occurred though, since $p_{1i_0} = 0$ means that Equation (34) of Definition 4 (with $n = 2$) is violated for sufficiently small $\epsilon > 0$. Note in particular that these conclusions about knowledge acquisition cannot be drawn from a purely Bayesian analysis.

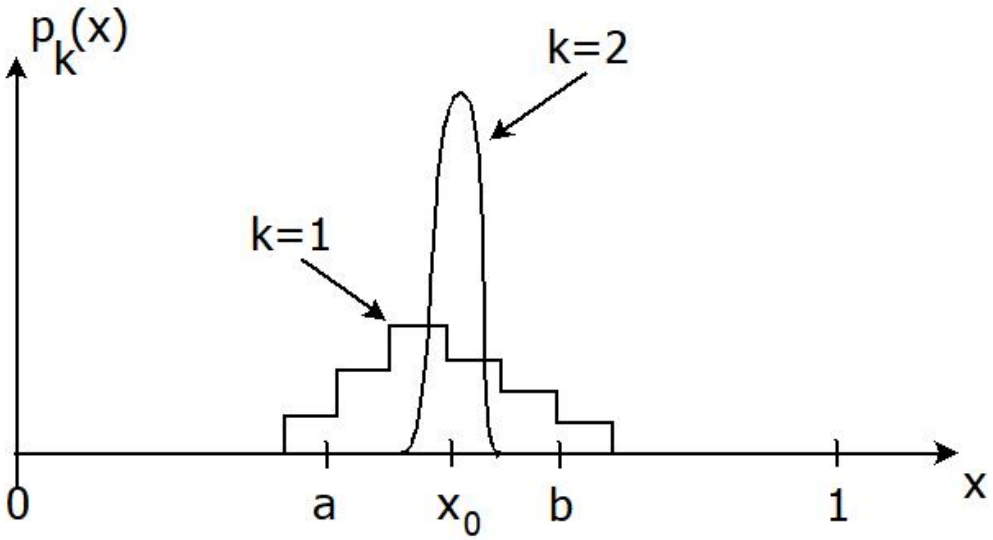

Figure 4. Posterior densities $p_1(x)$ and $p_2(x)$ after one and two steps of the discerment and data collection process of Example 2 when S is true ($x_0 \in [a,b]$). Since p_1 is measurable with respect to $\mathcal{G}_1$, it is piecewise-constant with step length δ. Note that strong learning and knowledge acquisition occurs.

Assume now that the contents of the book are not reliable. A probability measure P_2 on $[0,1]$ may be chosen so that it incorporates data Z_1 from the radiometric dating and data Z_2 from the book. This probability measure will also include information about the way the text of the book is believed to be unreliable. If the agent trusts Z_2 too much, it may happen that strong learning does not occur.

Example 3 (Future events). *A youth camp with certain outdoor activities is planned for a weekend. Let $\mathcal{X} = (0,1]^2$ denote the set of possible temperatures $x = (x_1, x_2)$ of the two days for which*

the camp is planned, each normalized within a range $0 \leq x_i \leq 1$. *The outdoor activities are only possible within a certain sub-range* $0 < a \leq x_1, x_2 \leq b < 1$ *of temperatures. The proposition*

$$S: \text{ it is possible to have the outdoor activities}$$

corresponds to a truth function $f(x) = \mathbf{1}\big(x \in [a,b]^2\big)$ *and* $A = [a,b]^2$. *The leaders have to travel to the camp five days before it starts and then make a decision on whether to bring equipment for the outdoor activities or for some other indoor activities. In the first step they consult weather forecast data* $D_1 = Z_1$, *with a* σ*-algebra* $\mathcal{G}_1$ *given by*

$$\sigma\{((i-1)\delta_1, i\delta_1] \times ((j-1)\delta_2, j\delta_2];\, 1 \leq i \leq N_1, 1 \leq j \leq N_2\},$$

which is $\sigma(Y_1)$, *the* σ*-algebra generated by* Y_1, *where* δ_1 *and* $\delta_2 > \delta_1$ *represent the maximal possible accuracy of weather forecasts five and six days ahead, respectively,* $N_i = 1/\delta_i$ *and* $Y_1(x) = ([x_1/\delta_1] + 1, [x_2/\delta_2] + 1)$. *Let* P_0 *be the uniform distribution on* $[0,1]^2$. *Due to* (3), P_1 *has a density function*

$$p_1(x) = N_1 N_2 \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} p_{1ij} \mathbf{1}\big((x_1, x_2) \in R_{ij}\big), \tag{68}$$

for some non-negative probability weights $p_{1ij} = p_{1i}(D_1) \geq 0$ *that sum to 1, with*

$$R_{ij} = ((i-1)\delta_1, i\delta_1] \times ((j-1)\delta_2, j\delta_2],$$

a rectangular region that corresponds to the ith temperature interval the first day of the camp and the jth temperature interval the second day. Consequently, the accuracy $\mathcal{G}_1$ *of weather forecast data forces* p_1 *to be constant over each* R_{ij}. *A* $\mathcal{G}_1$*-optimal measure assigns full weight 1 to the rectangle* R_{ij} *with* $i = [x_{01}/\delta_1] + 1$ *and* $j = [x_{02}/\delta_2] + 1$, *where* $x_0 = (x_{01}, x_{02})$ *represents the actual temperature the two days. Observe then that the* $\mathcal{G}_1$*-optimal measure is restricted to measurements that are accurate up to* δ_1 *and* δ_2; *therefore, it cannot do better than assigning the temperature to the intervals with sizes* δ_1 *and* δ_2 *to which the actual temperatures belong; however, it cannot say what the exact temperature is. The exact prediction requires an* $\mathcal{F}$*-optimal measure.*

In a second step, in order to get some additional information, the leaders of the camp consult a prophet. Let P_2 *refer to the probability measure based on the weather forecast* Z_1 *and the message* Z_2 *of the prophet, so that* $D_2 = (Z_1, Z_2)$ *and* $\mathcal{G}_2 = \mathcal{F}$. *If the prophet always speaks the truth, and if the leaders of the camp rely on his message, they will make use of the* $\mathcal{F}$*-optimal measure* $P_2 = \delta_{x_0}$, *corresponding weak (and full) learning, and a full amount of knowledge. In general, the camp leaders' prediction in step k is correct with*

$$\textit{probability} = \begin{cases} P_k\big([a,b]^2\big), & \text{if } x_0 \in [a,b]^2, \\ 1 - P_k\big([a,b]^2\big), & \text{if } x_0 \notin [a,b]^2. \end{cases}$$

If this probability is less than 1 for $k = 2$, *the reason is either that the prophet does not always speak the truth or that the leaders do not rely solely on the message of the prophet. In particular, it follows from Definition 3 that strong learning has occurred if*

$$\begin{aligned} 0 < I_1^+ = \log[P_1\big([a,b]^2\big)/(b-a)^2] < I_2^+ = \log[P_2\big([a,b]^2\big)/(b-a)^2], \quad \text{if } x_0 \in [a,b]^2, \\ 0 > I_1^+ = \log[P_1\big([a,b]^2\big)/(b-a)^2] > I_2^+ = \log[P_2\big([a,b]^2\big)/(b-a)^2], \quad \text{if } x_0 \notin [a,b]^2. \end{aligned} \tag{69}$$

Suppose the weather forecast and the message of the prophet are biased, but they still correctly predict whether outdoor activities are possible or not. Then, neither weak nor strong knowledge acquisition occurs, in spite of the fact that the strong learning condition (69) *holds. Note in particular that such a conclusion is not possible with a purely Bayesian analysis. Another scenario wherein neither (weak or strong) learning nor knowledge acquisition takes place is depicted in Figure 5.*

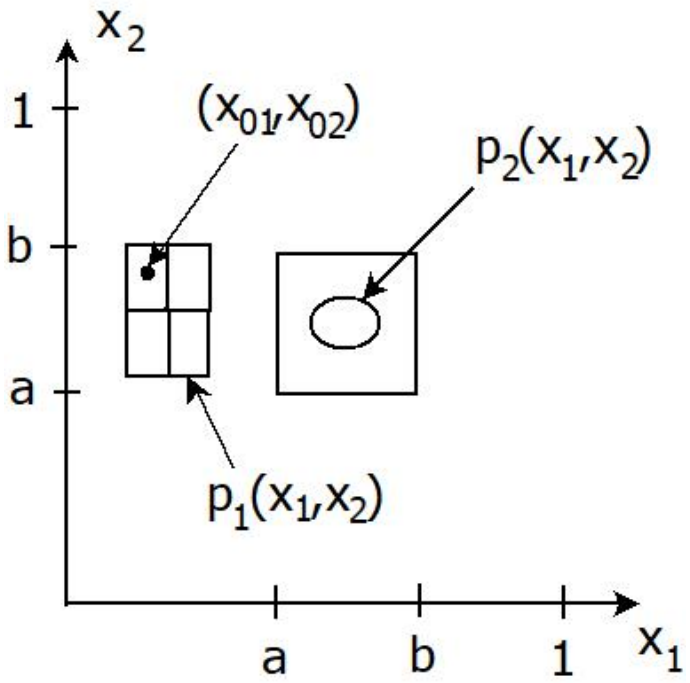

Figure 5. Posterior densities $p_1(x_1, x_2)$ and $p_2(x_1, x_2)$ after one and two steps of data collection for Example 3. Since $x_{01} < a$, it is not possible to have outdoor activities the first day of the camp. The weather forecast density p_1 is supported and piecewise-constant on the four rectangles with width δ_1 and height δ_2, corresponding to σ-algebra $\mathcal{G}_1$. The true temperatures (x_{01}, x_{02}) are within the support of p_1. On the other hand, the prophet incorrectly predicts that outdoor activities are possible both days; p_2 is supported on the ellipse. In this case, neither (weak or strong) learning nor knowledge acquisition takes place.

Example 4 (Replication of studies)**.** *Some researchers want to find the prevalence of the physical symptoms of a certain disease. Let $\mathcal{X} = [0,1]^2$ refer to the possible set of values $x = (x_1, x_2)$ for the prevalence of the symptoms, obtained from two different laboratories. The first value corresponds to the prevalence obtained in Laboratory 1, whereas the second value x_2 is obtained when Laboratory 2 tries to replicate the study of Laboratory 1. The board members of the company to which the two laboratories belong want to find out whether the two estimates are consistent, within some tolerance level $0 < \varepsilon < 1$. In that case, the second study is regarded as replication of the first one. The proposition*

$$S: \text{ the second study replicates the first one}$$

corresponds to a truth function $f(x) = \mathbf{1}(|x_2 - x_1| \leq \varepsilon)$ and

$$A = \{(x_1, x_2);\, |x_2 - x_1| \leq \varepsilon\}. \tag{70}$$

The true value $x_0 = (x_{01}, x_{02})$ represents the actual prevalences of the symptoms, obtained from the two laboratories under ideal conditions. Importantly, it may still be the case that $x_{01} \neq x_{02}$, if either the prevalence of the symptoms changes between the two studies and/or the two laboratories estimate the prevalences within two different subpopulations.

Let D_2 be a data set by which Laboratory 2 receives all needed data from Laboratory 1 in order to set up its study properly (so that, for instance, affection status is defined in the same way in the two laboratories). We will assume $Y_2(x_1, x_2) = (x_1, x_2)$, so that the corresponding σ-algebra

$$\mathcal{G}_2 = \mathcal{F} = \mathcal{B}(\mathcal{X}) = \mathcal{B}_0 \times \mathcal{B}_0,$$

corresponds to full discernment, with $\mathcal{B}(\mathcal{X})$ being the Borel σ-algebra on $\mathcal{X}$, whereas $\mathcal{B}_0 = \mathcal{B}([0,1])$ is the Borel σ-algebra on the unit interval (see Remark 1). If P_2 is the probability measure obtained from D_2, the probability of concluding that the second study replicated the first is

$$P_2(A) = \int_A dP_2(x_1, x_2). \tag{71}$$

In particular, when each laboratory makes use of data from all individuals in its subpopulation (which is either the same or not for the two laboratories), the $\mathcal{F}$-optimal probability measure (9) corresponds to

$$P_2 = \delta_{x_0} \Longrightarrow P_2(A) = \mathbf{1}(|x_{02} - x_{01}| \leq \varepsilon). \tag{72}$$

Now consider another scenario where Laboratory 2 only gets partial information from Laboratory 1. This corresponds to a data set D_1 with the same sampled individuals as in D_2, but Laboratory 2 has incomplete information from Laboratory 1 regarding the details of how the first study was set up. For this reason, they make use of a coarser σ-algebra, by which it is only possible to quantify prevalence with precision δ. If this σ-algebra is referred to as $\mathcal{B}_\delta \subset \mathcal{B}_0$, it follows that $Y_1(x_1, x_2) = (x_1, [x_2/\delta] + 1)$ and

$$\mathcal{G}_1 = \mathcal{B}_0 \times \mathcal{B}_\delta.$$

The corresponding loss of information is measured through a probability P_1 that has the same marginal distribution as P_2 for all events B that are discernible from $\mathcal{G}_1$, i.e.,

$$P_1(B) = P_2(B), \quad \forall B \in \mathcal{G}_1. \tag{73}$$

Hence, it follows from (30) and (73) that

$$dP_1(x_1, x_2) = N \sum_{j=1}^{N} p_{1j}(x_1) \mathbf{1}(x_2 \in R_j) dP_2(x_1),$$

where $N = 1/\delta$, $p_{1j}(x_1) = P_2(X_2 \in R_j \mid X_1 = x_1)$, and $R_j = ((j-1)\delta, j\delta]$ is the j-th possible region for the prevalence estimate of Laboratory 2. In particular, the probability that the second study replicates the first one is

$$P_1(A) = N \int_0^1 \sum_{j=1}^{N} p_{1j}(x_1) |R_j \cap [x_1 - \varepsilon, x_1 + \varepsilon]| dP_2(x_1). \tag{74}$$

If both laboratories perform a screening and collect data from all individuals in their regions, so that (72) holds, then P_1 is a $\mathcal{G}_1$-optimal measure according to (8), with

$$P_1(A) = N |R_{j_0} \cap [x_{01} - \varepsilon, x_{01} + \varepsilon]|, \tag{75}$$

and $j_0 = [x_{02}/\delta] + 1$. Making use of Definition 3, we notice that a sufficient condition for strong learning to occur is that P_0 has a uniform distribution on $\mathcal{X}$ (so that $P_0([a,b]^2) = 2\varepsilon - \varepsilon^2$), such that (72) and (75) hold, and

$$\begin{aligned} 0 < I_1^+ = \log[N|R_{j_0} \cap [x_{01} - \varepsilon, x_{01} + \varepsilon]|/(2\varepsilon - \varepsilon^2)] < I_2^+ = 1, \quad & \text{if } |x_{02} - x_{01}| < \varepsilon, \\ 0 > I_1^+ = \log[N|R_{j_0} \cap [x_{01} - \varepsilon, x_{01} + \varepsilon]|/(2\varepsilon - \varepsilon^2)] > I_2^+ = -\infty, \quad & \text{if } |x_{02} - x_{01}| > \varepsilon. \end{aligned}$$

With full information transfer between the two laboratories, the replication probabilities (71) and (72) based on data D_2 only depend on ε, whereas the corresponding replication probabilities (74) and (75) under incomplete information transfer between the laboratories and data D_1, also depending on δ. In particular, $P_1(A)$ will always be less than 1 when $2\varepsilon < \delta$, even when (75) holds and $x_{01} = x_{02}$. Moreover, δ sets the limit in terms of how much knowledge can be obtained from the two studies under incomplete information transfer, since

$$P_1(B_\epsilon(x_0)) < 1, \quad \text{for all } 0 < \epsilon < \delta.$$

Note that this last conclusion cannot be obtained from a Bayesian analysis, since a true pair x_0 of prevalences does not belong to a purely Bayesian framework.

Example 5 (Unmeasured confounding and causal inference)**.** *This example illustrates unmeasured confounding and causal inference. Let $q = n$ and $\mathcal{X} = \{0,1\}^n$. An individual is assigned a binary vector $x = (x_1, \ldots, x_n)$ of length n, where $x_n \in \{0,1\}$ codes for whether that person will have symptoms within five years ($x_n = 1$) or not ($x_n = 0$) that are associated with a certain mental disorder. The first component $x_1 \in \{0,1\}$ refers to the individual's binary exposure, whereas the other variables $x_k \in \{0,1\}$, $k = 2, \ldots, n-1$ are binary confounders. The truth function $f(x) = x_n$ corresponds to symptom status, whereas*

$$A = \{x \in \mathcal{X}; x_n = 1\}$$

represents the vectors x of all individuals in the population with symptoms. Consider the proposition

$$S: \text{ Adam will have the symptoms within five years,}$$

and let $x_0 = (x_{01}, \ldots, x_{0n})$ be the vector associated with Adam. We will introduce a sequence of probability measures $P_0, P_1, \ldots, P_n$, where P_0 represents the distribution of $X = (X_1, \ldots, X_n)$ in the whole population, whereas P_k corresponds to the conditional distribution of $X \sim P_0$, given that its first k covariates $D_k = (Z_1, \ldots, Z_k) = (x_{01}, \ldots, x_{0k}) \in \mathcal{D}_k = \{0,1\}^k$ have been observed, with values equal to those of Adam's first k covariates. Since the conditional distribution $D_k | x_0$ is non-random, it follows that

$$\bar{P}_k = P_k = \prod_{l=1}^{k} \delta_{x_{0l}} \times P_0\Big(\cdot \mid \{X_l = x_{0l}\}_{l=1}^{k} \Big) \tag{76}$$

for $k = 0, 1, \ldots, n-1$, whereas $\bar{P}_n = P_n = \delta_{x_0}$ for $k = n$. According to Definition 5, this implies that weak learning occurs with probability 1, and in particular that weak learning is expected to occur. If $Y_k(x_1, \ldots, x_n) = (x_1, \ldots, x_k)$, we have that

$$\mathcal{G}_k = 2^{\{0,1\}^k} \times \{0,1\}^{n-k} \tag{77}$$

for $k = 0, \ldots, n$. Note, in particular, that P_k is $\mathcal{G}_k$-optimal, corresponding to error-free measurement of Adam's first k covariates.

In order to specify the null distribution P_0, we assume that a logistic regression model [48]

$$\begin{aligned} P_0(X_n = 1 \mid x_1, \ldots, x_{n-1}) &= \frac{\exp\Big(\beta_0 + \sum_{k=1}^{n-1} \beta_k x_k\Big)}{1 + \exp\Big(\beta_0 + \sum_{k=1}^{n-1} \beta_k x_k\Big)} \\ &= g(x_1, \ldots, x_{n-1}) \end{aligned} \tag{78}$$

holds for the probability of having the symptoms within five years, conditionally on the $n-1$ covariates (one exposure and $n-2$ confounders). It is also assumed that the regression parameters $\beta_0, \ldots, \beta_{n-1}$ are known, so that g is known as well. It follows from Equations (76) and (78) that

$$\begin{aligned} P_k(A) &= P_0\Big(X_n = 1 \mid \{X_l = x_{0l}\}_{l=1}^{k}\Big) \\ &= E_{P_0}\Big[g(X_1, \ldots, X_{n-1}) \mid \{X_l = x_{0l}\}_{l=1}^{k}\Big] \\ &=: g_k(x_{01}, \ldots, x_{0k}) \end{aligned} \tag{79}$$

can be interpreted as increasingly better predictions of Adam's symptom status five years ahead, for $k = 0, 1, \ldots, n-1$, whereas $P_n(A) = f(x_0) = x_{0n}$ represents full knowledge of S. In particular, $P_0(A)$ is the prevalence of the symptoms in the whole population, whereas $P_1(A) = g_1(x_{01})$ is

Adam's predicted probability of having the symptoms when his exposure x_{01} is known, whereas none of his confounders are measured.

Suppose $x_2, \ldots, x_{n-1}$ are sufficient for confounding control, and that the exposure and the confounders (in principle) can be assigned. Let $x_0 = (x_{01}, \ldots, x_{0n})$ represent a hypothetical individual for which all covariates are assigned. Under a so called conditional exchangeability condition [16], it is possible to use a slightly different definition

$$\tilde{P}_k = \prod_{l=1}^{k} \delta_{x_{0l}} \times E_{P_0}\left[P_0\left(\cdot \mid \{X_l\}_{l=1}^{k}\right)\right]$$

of the probability measures in order to compute the counterfactual probability

$$\begin{aligned} h_k(x_{01}, \ldots, x_{0k}) &= \tilde{P}_k(A) \\ &= E_{P_0}[g(x_{01}, \ldots, x_{0k}, X_{k+1}, \ldots, X_{n-1})] \end{aligned}$$

of the potential outcome $X_n = 1$, under the scenario that the first k covariates were set to $x_{01}, \ldots, x_{0k}$. In particular, it is of interest to know how much the unknown causal risk ratio effect $h_1(1)/h_1(0)$ of the exposure maximally differs from the known risk ratio $g_1(1)/g_0(0)$ [49–52]. Note in particular that the corresponding logged quantities

$$\begin{array}{rcl} \log[g_1(1)/g_1(0)] & = & I_1^+(A;1) - I_1^+(A;0), \\ \log[h_1(1)/h_1(0)] & = & \tilde{I}_1^+(A;1) - \tilde{I}_1^+(A;0), \end{array}$$

can be expressed in terms of the active information

$$\begin{array}{rcl} I_1^+(A;x_{01}) & = & \log[P_1(A)/P_0(A)] \\ & = & \log[P_0(X_n = 1 \mid x_{01})/P_0(X_n = 1)], \\ \tilde{I}_1^+(A;x_{01}) & = & \log[\tilde{P}_1(A)/P_0(A)] \\ & = & \log[E_{P_0}(P_0(X_n = 1 \mid x_{01}, X_2, \ldots, X_{n-1}))/P_0(X_n = 1)]. \end{array}$$

9. Discussion

In this paper, we studied an agent's learning and knowledge acquisition within a mathematical framework of possible worlds. Learning is interpreted as an increased degree of true belief, whereas knowledge acquisition additionally requires that the belief is justified, corresponding to an increased belief in the correct world. The theory is put into a framework that involves elements of frequentism and Bayesianism, with possible worlds corresponding to the parameters of a statistical model, where only one parameter value is true, whereas the agent's beliefs are obtained from a posterior distribution. We formulated learning as a hypothesis test within this framework, whereas knowledge acquisition corresponds to consistency of posterior distributions. Importantly, we argue that a hybrid frequentist/Bayesian approach is needed in order to model mathematically the way in which philosophers distinguish learning from knowledge acquisition.

Some applications of our theory were provided in the examples of Section 8. Apart from those, we argue that our framework has quite general implications for machine learning, in particular, supervised learning. A typical task of machine learning is to obtain a predictor of a binary outcome variable $Y = f(x_0)$, when only incomplete information X of x_0 is obtained from training data. The performance of a machine learning algorithm is typically assessed in terms of prediction accuracy, that is, how well $f(X)$ approximates Y, with less focus on the closeness between X and x_0. In our terminology, the purpose of machine learning is learning rather than knowledge acquisition. This can often be a disadvantage, since knowledge acquisition often provides deeper insights than learning. For instance, full knowledge acquisition may fail asymptotically when $k \to \infty$, even when data are unbiased and interpreted correctly by the agent, if there is lacking discernment between the set of possible worlds $\mathcal{X}$, even in the limit $k \to \infty$.

On the other hand, it makes no sense to go beyond learning for game theory, where the purpose is to find the optimal strategy (an instance of knowledge-how). In more detail, let $x \in \mathcal{X} = \{0, \ldots, M-1\}$ refer to the strategy x of a player among a finite set of M possible strategies. The optimal strategy x_0 is the one that maximizes an expected reward function $R(x)$ for the actions taken with strategy x, D refers to data from previous games that a player makes use of to estimate $R(\cdot)$, and $\mathcal{G}$ represents the player's maximal possible discernment between strategies. Since the objective is to find the optimal strategy, it is natural to use a truth function

$$f(x) = \mathbf{1}(x = x_0), \tag{80}$$

with the associated set $A = A_0 = \{x_0\}$ of true worlds corresponding to the upper row of (1). It follows from Remark 7 that learning and knowledge acquisition are equivalent for game theory whenever (80) is used. Various algorithms, such as reinforcement learning [53] and sequential sampling models [54,55], could be used by a player in order to generate his beliefs P about which strategy is the best.

Many extensions of our work are possible. A first extension would be to generalize the framework of Theorem 1 and Example 1, where data $\{D_k = (Z_1, \ldots, Z_k)\}_{k=1}^n$ are collected sequentially according to a Markov process with increasing state space, without requiring that $\{Z_l\}_{l=1}^n$ are independent and identically distributed. We will mention two related models for which this framework applies. For both of these models a student's mastery of q skills (which represent knowledge how rather than knowledge that) is of interest. More specifically, $x = (x_1, \ldots, x_q)$ is a binary sequence of length q, with $x_i = 1$ or 0 depending on whether the student has acquired skill i or not, whereas D_k corresponds to exercises that are given to a student up to time k, and the student's answers to these exercises. It is also known which skills are required to solve each type of exercise. The first model is Bayesian knowledge tracing (BKT) [56], which has recently been analyzed using recurrent neural networks [1]. In BKT, a tutor trains the student to learn the q skills, so that the student's learning profile changes over time. At each time point, the tutor is free to choose the last exercises at time k based on previous exercises and what the student learnt up to time $k-1$. The goal of the tutoring is to reach a state $x_0 = (1, \ldots, 1)$ where the student has learned all skills. The most restrictive truth function (80) monitors whether the student has learned all skills or not, so that $P_k(A)$ is the probability that the student has learnt all skills at time k. In view of Remark 7, there is no distinction between learning and knowledge acquisition for such a truth function. A less restrictive truth function $f(x) = x_i$ focuses on whether the student has learnt skill i or not, so that $P_k(A)$ is the probability that the student learnt skill i at time k. The second model—the Bayesian version of Diagnostic Classification Models (DCMs) [57]—can be viewed as an extension of Illustration 1. The purpose of DCMs is not to train the student (as for knowledge tracing), but rather to diagnose the student's (or respondent's) current vector $x_0 = (x_{01}, \ldots, x_{0q})$, where $x_{0i} = 1$ or 0 if this particular student masters skill (or attribute) i or not. The exercises of DCM are usually referred to as items. Assume a truth function (80); $P_k(A)$ is the probability that the diagnostic test by time k has learnt which attributes the student masters. Note in particular that the student's attribute mastery profile x_0 is fixed, and it is rather the instructor that learns about x_0 when the student is being tested on new items.

A second extension would be to consider opinion making and consensus formation [58] for a whole group of N agents that are connected according to some social network. In this context, $\mathcal{G}_k$ represents the maximal amount of discernibility between possible worlds that is possible to achieve after k time steps based on external data (available to all agents) and information from other agents (which varies between agents and depends on properties of the social network). It is of interest in this context to study the dynamics of $\{P_{ki}(A)\}_{i=1}^N$ over time, where $P_{ki}(A)$ represents the belief of agent (or individual) i in proposition S after k time steps. This can be accomplished using a dynamical Bayesian network [59] with N nodes $i = 1, \ldots, N$ that represent individuals, associating each node i with a distribution P_{ki} over the set of possible worlds $\mathcal{X}$, corresponding to the beliefs of agent i at time k. A particularly interesting example in this vein would be to explore the

degree to which social media and social networks can influence learning and knowledge acquisition.

The third possible extension is related to consensus formation, but with a more explicit focus on how N decentralized agents make a collective decision. In order to illustrate this, we first describe a related model of cognition in bacterial populations. Marshall [60] has concluded that "the direction of causation in biology is cognition → code → chemicals". Cognition is observed when there is a discernment and data collection process that either optimizes code or improves the probability of a given chemical outcome. Accordingly, the strong learning process of Definition 3 can be used to model how biological cognition is attained (or at least is expected to be attained). For instance, in quorum sensing, once bacteria reach a critical density, they emit a chemical signal to ascertain the number of neighboring bacteria [61]; when a critical density is reached, the population performs certain functions as a unit (Table 1 of [62] presents several examples of bacterial functions partially controlled by quorum sensing). The proposition under consideration here is S: "the function is performed by at least a fraction ε of bacteria", where ε represents a critical density above which the bacteria act as a unit. The parameter $x = (x_1, \ldots, x_N)$ is a binary sequence reflecting the way in which a population of $N = q$ bacteria acts, so that $x_i = 1$ if bacterium i performs the function, whereas $f(x) = 1(x \in A) = 1(\sum_i x_i \geq \varepsilon N)$. For collective decisions, x_i rather represents a local decision of agent i, whereas $f(x)$ corresponds to the global decision of all agents. Learning about S at time $k = 1, 2, \ldots$ is described by $P_k(A)$, the probability that the population acts as a unit at time k. There is a phase transition at time k if the probabilities $P_1(A), \ldots, P_{k-1}(A)$ of the population acting as a unit are essentially null, whereas $P_k(A)$ becomes positive (and hence $I_k^+(A)$ gets large) when discernment ability and data are extended from $(\mathcal{G}_{k-1}, D_{k-1})$ to $(\mathcal{G}_k, D_k)$. This is closely related to the fine-tuning of biological systems [33,35] with f being a specificity function and A set of highly specified states, and fine-tuning after k steps of an algorithm that models the evolution of the system corresponding to $I_k^+(A)$ being large. As for the direction of causation from cognition to code, Kolmogorov's complexity, which measures the complexity of an outcome as the shortest code that produces it, can be used in place of or jointly with active information to measure learning [63].

A fourth theoretical extension is to consider the case $|\mathcal{X}| = \infty$. In this case, instead of the (discrete or continuous) uniform distribution given by (2), it will be necessary to consider more general maximum entropy distributions P_0, subject to some restrictions, in order to measure learning and knowledge acquisition [20,64–66].

A fifth extension is to consider models where the data sets D_k are not nested. This is of interest, for instance, in Example 5, when non-nested subsets of confounders are used to predict Adam's disease status. For such scenarios, it might be preferable to use information-based model selection criteria (such as maximizing AIN) in order to quantify learning [67], rather than sequentially testing various pairs of nested hypotheses by means of

$$\Delta I_k^+ = I_k^+ - I_{k-1}^+ = \log \frac{P_k(A)}{P_{k-1}(A)},$$

in order to assess whether learning has occurred in each step k (corresponding to strong learning of Definition 3).

A sixth extension would be to compare the proposed Bayesian/frequentist notions of learning and knowledge acquisition, with purely frequentist counterparts. Since learning corresponds to choosing between the two hypotheses in (18), we may consider a test that rejects the null hypothesis when the log likelihood ratio is small enough, or equivalently, when

$$\Lambda = -2 \log \frac{\max_{x \in A} L(D|x)}{\max_{x \in \mathcal{X}} L(D|x)} \geq t \tag{81}$$

for some appropriately chosen threshold t. The frequentist notion of learning is then formulated in terms of error probabilities of type I and II, analogously to (20), but for

the LR-test (81) rather than the Bayesian/frequentist test (19) with test statistic AIN, or the purely Bayesian approach that relies on posterior odds (21). A frequentist version of knowledge acquisition corresponds to using data D in order to produce a one-dimensional class of confidence regions CR for x_0, with a nominal coverage probability of CR that varies. In order to quantify how much knowledge that is acquired, it is possible to use the steepness of a curve that plots the actual coverage probability $P(x_0 \in \text{CR})$ as a function of the volume $|\text{CR}|$. However, a disadvantage of the frequentist versions of learning and knowledge acquisition is that they do not involve degrees of beliefs, the philosophical starting point of this article. This is related to the critique of frequentist hypothesis testing offered in [68]. Since no prior probabilities are allowed, within a frequentist setting, important notions such as the false report probability (FRP) and true report probability (TRP) are not computable, leading to many non-replicated findings.

A seventh extension is to consider multiple propositions $S_1, \ldots, S_m$, as in [69,70]. For each possible world $x \in \mathcal{X}$, we let $f : \mathcal{X} \to \{0,1\}^m$ be a truth function such that $f(x) = (f_1(x), \ldots, f_m(x))$, with $f_i(x) = 1$ (0) if S_i is true (false) in world x. It is then of interest to develop a theory of learning and knowledge acquisition of these m propositions. To this end, for each $y = (y_1, \ldots, y_m) \in \{0,1\}^m$, let $A_y = \{x \in \mathcal{X}; f(x) = y\}$ refer to the set of worlds for which the truth value of S_i is y_i for $i = 1, \ldots, m$. Learning is then a matter of determining which A_y is true (the one for which $x_0 \in A_y$), whereas justified true beliefs in $S_1, \ldots, S_m$ amount to finding x_0 as well. Learning of statements such as $S_i \vee S_j$ and $S_i \wedge S_i$ can be addressed using the $m = 1$ theory of this paper, since they correspond to binary-valued truth functions $\tilde{f}(x) = f_i(x) + f_j(x) - f_i(x)f_j(x)$ and $\tilde{f}(x) = f_i(x)f_j(x)$, respectively.

Author Contributions: Conceptualization and methodology: O.H., D.A.D.-P., J.S.R.; writing—original draft preparation: O.H.; writing—review and editing: D.A.D.-P., J.S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors wish to thank Glauco Amigo at Baylor University for his help with producing Figure 3. We also appreciate the comments of three anonymous reviewers that made it possible to considerably improve the quality of the paper.

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

References

1. Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.J.; Sohl-Dickstein, J. Deep Knowledge Tracing. In Proceedings of the Neural Information Processing Systems (NIPS) 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 505–513.
2. Pavese, C. Knowledge How. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2021.
3. Agliari, E.; Pachón, A.; Rodríguez, P.M.; Tavani, F. Phase transition for the Maki-Thompson rumour model on a small-world network. *J. Stat. Phys.* **2017**, *169*, 846–875. [CrossRef]
4. Lyons, R.; Peres, Y. *Probability on Trees and Networks*; Cambridge University Press: Cambridge, UK, 2016.
5. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **1998**, *393*, 440–442. [CrossRef]
6. Embreston, S.E.; Reise, S.P. *Item Response Theory for Psychologists*; Psychology Press: New York, NY, USA, 2000.
7. Stevens, S.S. On the Theory of Scales of Measurement. *Science* **1946**, *103*, 677–680. [CrossRef] [PubMed]
8. Thompson, B. *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*; American Psychological Association: Washington, DC, USA, 2004.
9. Gettier, E.L. Is Justified True Belief Knowledge? *Analysis* **1963**, *23*, 121–123. [CrossRef]
10. Ichikawa, J.J.; Steup, M. The Analysis of Knowledge. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2018.
11. Hájek, A. Probability, Logic, and Probability Logic. In *The Blackwell Guide to Philosophical Logic*; Goble, L., Ed.; Blackwell: Hoboken, NJ, USA, 2001; Chapter 16, pp. 362–384.
12. Demey, L.; Kooi, B.; Sack, J. Logic and Probability. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2019.
13. Hájek, A. Interpretations of Probability. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2019.

14. Savage, L. *The Foundations of Statistics*; Wiley: Hoboken, NJ, USA, 1954.
15. Swinburne, R. *Epistemic Justification*; Oxford University Press: Oxford, UK, 2001.
16. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
17. Berger, J. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed.; Springer: New York, NY, USA, 2010.
18. Dembski, W.A.; Marks, R.J., II. Bernoulli's Principle of Insufficient Reason and Conservation of Information in Computer Search. In Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, TX, USA, 11–14 October 2009; pp. 2647–2652. [CrossRef]
19. Dembski, W.A.; Marks, R.J., II. Conservation of Information in Search: Measuring the Cost of Success. *IEEE Trans. Syst. Man Cybern.-Part A Syst. Hum.* **2009**, *5*, 1051–1061. [CrossRef]
20. Díaz-Pachón, D.A.; Marks, R.J., II. Generalized active information: Extensions to unbounded domains. *BIO-Complexity* **2020**, *2020*, 1–6. [CrossRef]
21. Shafer, G. Belief functions and parametric models. *J. R. Stat. Soc. Ser. B* **1982**, *44*, 322–352. [CrossRef]
22. Wasserman, L. Prior envelopes based on belief functions. *Ann. Stat.* **1990**, *18*, 454–464. [CrossRef]
23. Dubois, D.; Prade, H. Belief functions and parametric models. *Int. J. Approx. Reason.* **1992**, *6*, 295–319. [CrossRef]
24. Denoeux, T. Decision-making with belief functions: A review. *Int. J. Approx. Reason.* **2019**, *109*, 87–110.
25. Hopkins, E. Two competing models of how people learn in games. *Econometrica* **2002**, *70*, 2141–2166. [CrossRef]
26. Stoica, G.; Strack, B. Acquired knowledge as a stochastic process. *Surv. Math. Appl.* **2017**, *12*, 65–70.
27. Taylor, C.M. A Mathematical Model for Knowledge Acquisition. Ph.D. Thesis, University of Virginia, Charlottesville, VA, USA, 2002.
28. Popper, K. *The Logic of Scientific Discovery*; Hutchinson: London, UK, 1968.
29. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241. [CrossRef]
30. Hössjer, O. Modeling decision in a temporal context: Analysis of a famous example suggested by Blaise Pascal. In *The Metaphysics of Time, Themes from Prior. Logic and Philosophy of Time*; Hasle, P., Jakobsen, D., Øhrstrøm, P., Eds.; Aalborg University Press: Aalborg, Denmark, 2020; Volume 4, pp. 427–453.
31. Kowner, R. Nicholas II and the Japanese body: Images and decision-making on the eve of the Russo-Japanese War. *Psychohist. Rev.* **1998**, *26*, 211–252.
32. Hössjer, O.; Díaz-Pachón, D.A.; Chen, Z.; Rao, J.S. Active information, missing data, and prevalence estimation. *arXiv* **2022**, arXiv:2206.05120.
33. Díaz-Pachón, D.A.; Hössjer, O. Assessing, testing and estimating the amount of fine-tuning by means of active information. *Entropy* **2022**, *24*, 1323. [CrossRef]
34. Szostak, J.W. Functional information: Molecular messages. *Nature* **2003**, *423*, 689. [CrossRef] [PubMed]
35. Thorvaldsen, S.; Hössjer, O. Using statistical methods to model the fine-tuning of molecular machines and systems. *J. Theor. Biol.* **2020**, *501*, 110352. [CrossRef] [PubMed]
36. Díaz-Pachón, D.A.; Sáenz, J.P.; Rao, J.S. Hypothesis testing with active information. *Stat. Probab. Lett.* **2020**, *161*, 108742. [CrossRef]
37. Montañez, G.D. A Unified Model of Complex Specified Information. *BIO-Complexity* **2018**, *2018*, 1–26. [CrossRef]
38. Yik, W.; Serafini, L.; Lindsey, T.; Montañez, G.D. Identifying Bias in Data Using Two-Distribution Hypothesis Tests. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, UK, 19–21 May 2021; ACM: New York, NY, USA, 2022; pp. 831–844. [CrossRef]
39. Kallenberg, O. *Foundations of Modern Probability*, 3rd ed.; Springer: New York, NY, USA, 2021; Volume 1.
40. Ghosal, S.; van der Vaart, A. *Fundamentals of Nonparametric Bayesian Inference*; Cambridge University Press: Cambridge, UK, 2017.
41. Shen, W.; Tokdar, S.T.; Ghosal, S. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **2013**, *100*, 623–640. [CrossRef]
42. Barron, A.R. Uniformly Powerful Goodness of Fit Tests. *Ann. Stat.* **1989**, *17*, 107–124. [CrossRef]
43. Freedman, D.A. On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *Ann. Math. Stat.* **1963**, *34*, 1386–1403. [CrossRef]
44. Cam, L.L. Convergence of Estimates Under Dimensionality Restrictions. *Ann. Stat.* **1973**, *1*, 38–53. [CrossRef]
45. Schwartz, L. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie Verw Geb.* **1965**, *4*, 10–26. [CrossRef]
46. Cam, L.L. *Asymptotic Methods in Statistical Decision Theory*; Springer: New York, NY, USA, 1986.
47. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: New York, NY, USA, 1998.
48. Agresti, A. *Categorical Data Analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
49. Robins, J.M. The analysis of Randomized and Nonrandomized AIDS Treatment Trials Using A New Approach to Causal Inference in Longitudinal Studies. In *Health Service Research Methodology: A Focus on AIDS*; Sechrest, L., Freeman, H., Mulley, A., Eds.; U.S. Public Health Service, National Center for Health Services Research: Washington, DC, USA, 1989; pp. 113–159.
50. Manski, C.F. Nonparametric Bounds on Treatment Effects. *Am. Econ. Rev.* **1990**, *80*, 319–323.
51. Ding, P.; VanderWeele, T.J. Sensitivity Analysis Without Assumptions. *Epidemilogy* **2016**, *27*, 368–377. [CrossRef]
52. Sjölander, A.; Hössjer, O. Novel bounds for causal effects based on sensitivity parameters on the risk difference scale. *J. Causal Inference* **2021**, *9*, 190–210. [CrossRef]
53. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.

54. Ratcliff, R.; Smith, P.L. A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychol. Rev.* **2004**, *111*, 333–367. [CrossRef]
55. Chen, W.J.; Krajbich, I. Computational modeling of epiphany learning. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 4637–4642. [CrossRef] [PubMed]
56. Corbett, A.T.; Anderson, J.R. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Model. User-Adapt. Interact.* **1995**, *4*, 253–278. [CrossRef]
57. Oka, M.; Okada, K. Assessing the Performance of Diagnostic Classification Models in Small Sample Contexts with Different Estimation Methods. *arXiv* **2022**, arXiv:2104.10975.
58. Hirscher, T. Consensus Formation in the Deffuant Model. Ph.D. Thesis, Division of Mathematics, Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden, 2014.
59. Murphy, K.P. Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. Thesis, University of California, Berkeley, CA, USA, 2002.
60. Marshall, P. Biology transcends the limits of computation. *Prog. Biophys. Mol. Biol.* **2021**, *165*, 88–101. [CrossRef] [PubMed]
61. Atkinson, S.; Williams, P. Quorum sensing and social networking in the microbial world. *J. R. Soc. Interface* **2009**, *6*, 959–978. [CrossRef] [PubMed]
62. Shapiro, J.A. All living cells are cognitive. *Biochem. Biophys. Res. Commun.* **2020**, *564*, 134–149. [CrossRef]
63. Ewert, W.; Dembski, W.; Marks, R.J., II. Algorithmic Specified Complexity in the Game of Life. *IEEE Trans. Syst. Man Cybern. Syst.* **2015**, *45*, 584–594. [CrossRef]
64. Díaz-Pachón, D.A.; Hössjer, O.; Marks, R.J., II. Is Cosmological Tuning Fine or Coarse? *J. Cosmol. Astropart. Phys.* **2021**, *2021*, 020. [CrossRef]
65. Díaz-Pachón, D.A.; Hössjer, O.; Marks, R.J., II. Sometimes size does not matter. *arXiv* **2022**, arXiv:2204.11780.
66. Zhao, X.; Plata, G.; Dixit, P.D. SiGMoiD: A super-statistical generative model for binary dataP. *PLoS Comput. Biol.* **2021**, *17*, e1009275. [CrossRef]
67. Stephens, P.A.; Buskirk, S.W.; Hayward, G.D.; del Río, C.M. Information theory and hypothesis testing: A call for pluralism. *J. Appl. Ecol.* **2005**, *42*, 4–12. [CrossRef]
68. Szucs, D.; Ioannidis, J.P.A. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Front. Hum. Neurosci.* **2017**, *11*, 390. [CrossRef] [PubMed]
69. Cox, R.T. *The Algebra of Probable Inference*; Johns Hopkins University Press: Baltimore, MD, USA, 1961.
70. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003. [CrossRef]