




STATISTICAL LEARNING DOES NOT ALWAYS ENTAIL KNOWLEDGE

BY DANIEL ANDRÉS DÍAZ-PACHÓN^{1,a} , H. RENATA GALLEGOS^{1,b},
OLA HÖSSJER^{2,c} , AND J. SUNIL RAO^{3,d} ,

¹*Division of Biostatistics, University of Miami, ddiaz3@miami.edu; h.gallegos@med.miami.edu*

²*Department of Mathematics, Stockholm University, ola@math.su.se*

³*Department of Biostatistics, University of Minnesota, js-rao@umn.edu*

In this paper, we study learning and knowledge acquisition (LKA) of an agent about a proposition that is either true or false. We use a Bayesian approach, where the agent receives data to update his beliefs about the proposition according to a posterior distribution. The LKA is formulated in terms of active information, with data representing external or exogenous information that modifies the agent’s beliefs. It is assumed that data provide details about a number of features that are relevant to the proposition. We show that this leads to a Gibbs distribution posterior, which is in maximum entropy relative to the prior, conditioned on the side constraints that the data provide in terms of the features. We demonstrate that full learning is sometimes not possible and full knowledge acquisition is never possible when the number of extracted features is too small. We also distinguish between primary learning (receiving data about features of relevance for the proposition) and secondary learning (receiving data about the learning of another agent). We argue that this type of secondary learning does not represent true knowledge acquisition. Our results have implications for statistical learning algorithms, and we claim that such algorithms do not always generate true knowledge. The theory is illustrated with several examples.

1. Introduction.

1.1. *Learning and knowledge acquisition.* In the current era of scientific computing, when large language models have seemingly achieved surprising levels of understanding and discussions about artificial general intelligence are as abundant as nebulous, proper definitions that can be accurately quantified are conspicuous by their absence. For instance, what do we mean by “understanding” and “intelligence” in the previous paragraph? If explainable AI is going to explain anything, it does require clear concepts capable of guiding the discussion to reach valid conclusions. Philosophers usually define knowledge as “justified true belief” [21, 25, 37]. This means that an agent \mathcal{A} *knows* a proposition p if the following three properties are satisfied:

LK1 \mathcal{A} believes p ,

LK2 p is true,

LK3 \mathcal{A} ’s belief about p is justified.

If only properties **LK1** and **LK2** are satisfied, \mathcal{A} *learns* p . Clearly, acquiring knowledge requires more than learning. Therefore, even before further theoretical developments, we obtain a simple but revealing fact:

CLAIM 1. *Statistical learning does not always entail knowledge.*

MSC2020 subject classifications: Primary 60A99, 62A01; secondary 68T01, 62B10.

Keywords and phrases: active information, discernment, Gibbs distribution.

The mathematical formulation of learning and knowledge acquisition, based on **LK1-LK3**, was introduced in [24]. The main idea is that agent \mathcal{A} uses data D to learn and acquire knowledge about p . This approach has already been applied to determine which cases of cosmological fine-tuning can be known [15] (see also [13, 14]). Our approach to learning and knowledge acquisition goes further in four ways:

- (i) We develop the notion of discernment that was introduced in [24], further quantifying how it imposes limits on learning and knowledge acquisition.
- (ii) We focus on learning through feature extraction and Gibbs distributions, which is a natural and powerful approach.
- (iii) We motivate Claim 1 through multiple examples and results in which knowledge cannot be acquired, even if *full* learning is attained. This point was suggested but not fully elaborated in [24].
- (iv) We introduce the concepts of primary and secondary learning.

1.2. Active information. In order to describe (i)-(iv) in more detail, we will introduce local measures of information since our approach to learning and knowledge acquisition depends fundamentally on such measures. This is interesting because Shannon’s information theory has been almost exclusively focused on global averages such as entropy, mutual information, relative entropy, etc. However, recent decades have seen a resurgence of unaveraged measures of information like local active information storage and local transfer entropy. These measures have been used in origin of life [8, 40, 42], neuroscience [41, 43] as well as cancer research and cell communication [30, 31]. All such measures can be seen as mathematical extensions of the more basic active information, which was originally proposed to measure the amount of exogenous information infused by a programmer in a search, compared to the endogenous information generated by a blind search [9, 10]. Formally, if the distributions of the outcome of the programmer and the blind search are represented by two probability measures \mathbf{P} and \mathbf{P}_0 defined on the same measurable space $(\mathcal{X}, \mathcal{F})$, active information for a specific target $T \subset \mathcal{X}$ is defined as

$$(1) \quad I^+(T) = I^+(T; \mathbf{P}_0, \mathbf{P}) = \log \frac{\mathbf{P}(T)}{\mathbf{P}_0(T)},$$

where we assume $0/0 = 0$ by continuity. In particular, if the programmer reaches the target with certainty ($\mathbf{P}(T) = 1$), then (1) reduces to the self-information of T .

To this point, active information has been used in several areas. For instance, in genetics, to quantify functional information in genetic sequence data [38, 39], and to compare selectively non-neutral models to neutral ones in population genetics, where T was the event that a given allele gets fixed [16]; in bump-hunting, using machine learning algorithms to find a bump T [19, 29]; and in decision theory, to construct hypothesis tests that quantify the amount of information added, or needed, to produce an event T [12, 18].

1.3. A mixed frequentist-Bayesian framework for learning and knowledge acquisition. Following [24], in this article we apply active information to formalize the concepts **LK1-LK3** behind learning and knowledge acquisition. To this end, it is assumed that \mathcal{X} is a set of parameters of a statistical model; in this context, we take a mixed frequentist and Bayesian approach. On the one hand, it is postulated that one element $x_0 \in \mathcal{X}$ is the true parameter value (a frequentist assumption). On the other hand, uncertainty about x_0 is formulated as a probability measure on \mathcal{X} that varies between persons (a Bayesian assumption). More specifically, \mathbf{P} and \mathbf{P}_0 represent degrees of beliefs about $x_0 \in \mathcal{X}$, of an agent \mathcal{A} and an ignorant person \mathcal{I} , respectively. It is assumed that \mathcal{A} acquired data D that \mathcal{I} lacks, so that \mathbf{P} and \mathbf{P}_0 are posterior and prior distributions on \mathcal{X} that represent degrees of beliefs of \mathcal{A} about x_0 , after

and before he received data. In particular, if we choose T as the set of parameter values for which a given proposition p is true, then the objective of \mathcal{A} is to use data to learn whether the proposition is true ($x_0 \in T$) or not ($x_0 \notin T$), as quantified by the active information $I^+(T)$ in (1). In this case, data represents the exogenous information that helps agent \mathcal{A} modify his beliefs **LK1** about T compared to the ignorant person \mathcal{I} . Knowledge acquisition goes beyond learning since it additionally requires **LK3** that \mathcal{A} learns about the proposition for the right reason. This corresponds to increasingly correct beliefs about x_0 , not only increasingly correct beliefs of whether $x_0 \in T$ or not (as for learning).

Our approach proposes a very sensible solution to the old dispute between Bayesians and frequentists. We consider propositions and states of reality that are objectively true or false, but learning and knowledge are naturally Bayesian. Thus, ontology is partially frequentist, whereas epistemology is Bayesian. Our definitions differentiate between them, and this is an essential aspect of our theory.

1.4. The novelties of this article. Given the mathematical framework outlined in Section 1.3, the novelties (i)-(iv) for learning and knowledge acquisition mentioned above can be phrased as follows. Starting with (i), discernment is a crucial aspect of \mathcal{A} 's learning and knowledge acquisition process, which quantifies his ability to separate elements of \mathcal{X} from each other. We assume that \mathcal{A} 's discernment is larger than that of the ignorant person \mathcal{I} . We mean by this that \mathcal{A} 's beliefs \mathbf{P} are measurable with respect to a finer σ -field on \mathcal{X} than the beliefs \mathbf{P}_0 of \mathcal{I} . We prove general results on how \mathcal{A} 's σ -field affects his potential to learn and acquire knowledge.

As for (ii), we assume that data provide agent \mathcal{A} with details about (modifies his beliefs in) the values of a number of features of relevance for learning proposition p . Then \mathcal{A} forms his likelihood in such a way that \mathbf{P} maximizes entropy relative to \mathbf{P}_0 , among all probability measures on \mathcal{X} that are consistent with the beliefs of \mathcal{A} about the values of the features. As we shall see, this implies that \mathbf{P} belongs to a family of Gibbs distributions.

Novelty (ii) also has relevance for (iii) since feature extraction is a commonly used technique for data reduction within statistical learning (see, e.g., [22, Section 5.3]). But, as a consequence of the data processing inequality, feature extraction potentially implies a loss of information, regardless of how large the data set used to form \mathcal{A} 's beliefs about the values of the features is [7, Section 2.8], [11, Problem 2.1]. This implies that the Gibbs distribution beliefs of \mathcal{A} about the value of x_0 are limited by which features are selected in the first place. We give a number of examples of how this provides fundamental limits in terms of learning and knowledge acquisition.

The concept of secondary learning (iv) refers to the learning process of another agent $\tilde{\mathcal{A}}$ who lacks data D but, on the other hand, uses other data \tilde{D} to learn how much \mathcal{A} learned and acquired knowledge about the proposition. In other words, $\tilde{\mathcal{A}}$ learns and acquires knowledge about \mathcal{A} 's learning, but not necessarily about the proposition p itself. This also has an impact on (iii) since machine learning algorithms often recapitulate the beliefs of humans, thereby performing secondary (rather than primary) learning and knowledge acquisition. We also demonstrate that the long-term effects of secondary learning are very similar to those of synthetic primary learning, whereby a third agent \mathcal{A}' learns from synthetic primary data D' generated by \mathcal{A} .

1.5. Organization of article. Our paper is organized as follows. In Section 2, we define what it means that agent \mathcal{A} has learned whether a proposition is true or not and whether he acquired knowledge about the proposition or not. Then, in Section 3, we introduce a general framework for choosing the posterior distribution \mathbf{P} as a Gibbs distribution that maximizes the entropy relative to \mathbf{P}_0 , given side constraints that data D provide. The concepts of Sections 2 and 3 are applied to learning and knowledge acquisition for feature-like data and

Gibbs distributions in Section 4 and to secondary learning in Section 5. A discussion is provided in Section 6, whereas mathematical proofs are provided in Section 7.

2. Learning and knowledge. In this section, we reproduce the definitions of learning and knowledge acquisition in [24]. We also elaborate more on the concepts of σ -fields and discernment. In this context, we prove some new results (Proposition 2.1 and Theorem 2.4).

In order to formalize the notions of learning and knowledge acquisition, suppose we have a set of possible worlds defined by the space of parameters \mathcal{X} (i.e., each parameter value $x \in \mathcal{X}$ defines a world), where x_0 represents the true world, whereas $\{x_0\}^c = \mathcal{X} \setminus \{x_0\}$ is a collection of counterfactuals. For a given proposition p , we define a truth function $f_p : \mathcal{X} \rightarrow \{0, 1\}$ s.t.

$$(2) \quad f_p(x) = \begin{cases} 1 & \text{if } p \text{ is true in the world } x, \\ 0 & \text{if } p \text{ is false in the world } x. \end{cases}$$

Our goal is to learn $f_p(x_0)$, the truth value of the proposition in the true world. To accomplish this, we define the set

$$(3) \quad \mathbb{T} = \{x \in \mathcal{X} : f_p(x) = 1\}$$

of worlds in which proposition p is true. The fact that p is either true or false in the true world (i.e., $f_p(x_0) \in \{0, 1\}$) is an assumption on the nature of reality aligned with a frequentist understanding of $f_p(x_0)$.

2.1. Discernment and belief. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space that we define in the largest possible generality. That is, we assume that $\mathcal{F} = \sigma(\mathcal{O})$ is the Borel σ -field for the collection \mathcal{O} of open sets of \mathcal{X} that makes $(\mathcal{X}, \mathcal{O})$ a topological space. An agent \mathcal{A} will assign its belief about x_0 according to a probability measure \mathbf{P} , whereas an ignorant agent \mathcal{I} will assign its belief about x_0 following a probability measure \mathbf{P}_0 . Thus, \mathbf{P} and \mathbf{P}_0 are the respective predictors of \mathcal{A} and \mathcal{I} for x_0 , the value of the true world. With a slight abuse of notation, we refer to $\mathbf{P}_0(x)$ and $\mathbf{P}(x)$ as densities, regardless of whether the corresponding probability measures are absolutely continuous, discrete or a mixture of both. For each set $A \in \mathcal{F}$, agents \mathcal{I} and \mathcal{A} will assign probabilities to A by integrating over A their density functions $\mathbf{P}_0(x)$ and $\mathbf{P}(x)$. More explicitly, the beliefs of \mathcal{A} about A , in the presence of some data

$$(4) \quad D \in \Delta$$

that \mathcal{I} does not possess, are obtained as

$$(5) \quad \mathbf{P}(A) = \int_A \mathbf{P}(x) dx = \frac{\mathbf{L}(D | A) \mathbf{P}_0(A)}{\mathbf{L}(D)},$$

where dx is the Lebesgue measure $\nu(dx)$ of a Euclidean space if \mathcal{X} is an open subset this space and \mathbf{P} is absolutely continuous, whereas dx is the counting measure if \mathcal{X} is finite or countable. Moreover, $\mathbf{L}(D | A) = \int_A \mathbf{L}(D|x) \mathbf{P}_0(x) dx / \mathbf{P}_0(A)$ is the average likelihood of the parameters $x \in A$ given the data D , whereas $\mathbf{L}(D)$ quantifies the overall strength of evidence D , from the perspective of \mathcal{A} . In more detail, we assume that there is a random variable D taking values on some measurable space (Δ, \mathcal{D}) . For some underlying sample space Ω , we define the random element $(X, D) : \Omega \rightarrow \mathcal{X} \times \Delta$ that is $(\mathcal{F} \times \mathcal{D})$ -measurable. Moreover, to the measurable product space $(\mathcal{X} \times \Delta, \mathcal{F} \times \mathcal{D})$ we associate a joint law $\mathbf{Q}^*(x, \delta) = \mathbf{P}_0(x) \mathbf{L}(\delta | x)$ with marginals

$$(6) \quad \int_{\mathcal{X}} \mathbf{Q}^*(dx, \delta) = \mathbf{L}(\delta), \quad \int_{\Delta} \mathbf{Q}^*(x, d\delta) = \mathbf{P}_0(x).$$

Thus, the beliefs of \mathcal{I} correspond to the density of X , whereas the posterior beliefs of agent \mathcal{A} are obtained as the conditional density of X given the event $\{D = D\}$, expressed as

$$(7) \quad \mathbf{P}(x) := \mathbf{Q}^*(x | D) = \frac{\mathbf{Q}^*(x, D)}{\int_{\mathcal{X}} \mathbf{Q}^*(dx, D)}.$$

The densities \mathbf{P}_0 and \mathbf{P} are measurable with respect to σ -fields $\mathcal{G}_{\mathcal{I}}$ and $\mathcal{G}_{\mathcal{A}}$, respectively, with $\mathcal{G}_{\mathcal{I}} \subset \mathcal{G}_{\mathcal{A}} \subset \mathcal{F}$. This means that the beliefs of \mathcal{A} and \mathcal{I} are restricted to the information in $\mathcal{G}_{\mathcal{A}}$ and $\mathcal{G}_{\mathcal{I}}$, respectively. If

$$(8) \quad \mathcal{G}_{\mathcal{A}} = \sigma(A_1, A_2, \dots)$$

is generated by a countable partition

$$(9) \quad \mathcal{P} = \{A_1, A_2, \dots\}$$

of \mathcal{X} , it is assumed that

$$(10) \quad \mathbf{P}(x) = \sum_i p_i \mathbb{1}_{A_i}(x)$$

is piecewise constant over, and hence measurable with respect to, the sets in \mathcal{P} that generate $\mathcal{G}_{\mathcal{A}}$. Similarly, the density of \mathbf{P}_0 is piecewise constant over the sets of a partition \mathcal{P}_0 that is coarser than (9). The assumption that agent \mathcal{A} is able to discern from a finer partition \mathcal{P} of \mathcal{X} is natural, as it is often the case with refined experiments that they induce finer σ -fields for the potential resolution that data D can provide about $x \in \mathcal{X}$. This is particularly obvious in the most extreme case, when the ignorant agent's discernment is the trivial σ -field on \mathcal{X} , generated by a partition $\mathcal{P}_0 = \{\mathcal{X}\}$, and given by

$$(11) \quad \mathcal{G}_{\mathcal{I}} = \{\mathcal{X}, \emptyset\}.$$

In particular, if (11) holds and \mathcal{X} is bounded, then \mathbf{P}_0 has a constant density function over \mathcal{X} , making it necessarily the uniform distribution

$$(12) \quad \mathbf{P}_0(A) = \frac{|A|}{|\mathcal{X}|}$$

for all $A \in \mathcal{F}$, where $|\mathcal{X}|$ refers to the number of elements of \mathcal{X} for a finite set, or the Lebesgue measure $|\mathcal{X}| = \nu(\mathcal{X})$ when \mathcal{X} is a bounded subset of Euclidean space.

It follows from equation (8) that \mathcal{A} has no advantage over \mathcal{I} in terms of discerning how the probability mass is distributed *inside* the sets A_i that generate $\mathcal{G}_{\mathcal{A}}$. On the other hand, if $\mathcal{G}_{\mathcal{A}} = \mathcal{F}$, there is maximum flexibility in the choice of \mathbf{P} . Therefore, it follows that the σ -fields generated by countable partitions of the space \mathcal{X} represent how much \mathcal{A} and \mathcal{I} are maximally able to discern different possible worlds in \mathcal{X} . We formalize this with the following definition.

DEFINITION 2.1 (Discernment). Let $\mathcal{G}_{\mathcal{A}}$ be generated by a countable partition of \mathcal{X} , as in (8). For any σ -field \mathcal{G} such that $\mathcal{G}_{\mathcal{I}} \subset \mathcal{G}_{\mathcal{A}} \subset \mathcal{G} \subset \mathcal{F}$, and any \mathcal{F} -measurable function g ,

$$(13) \quad \mathbf{E}_{\mathbf{P}}(g \parallel \mathcal{G}) = \mathbf{E}_{\mathbf{P}_0}(g \parallel \mathcal{G}) \quad \text{a.s.}$$

That is, the conditional expectation function $x \mapsto \mathbf{E}_{\mathbf{P}}(g(X) \parallel \mathcal{G})(x)$ of agent \mathcal{A} is the same as that of the ignorant agent \mathcal{I} . In particular, if $g(x) = \mathbb{1}_A(x) = \mathbb{1}\{x \in A\}$, the indicator function of $A \subset \mathcal{X}$, then $\mathbf{E}_{\mathbf{P}}(\mathbb{1}_A \parallel \mathcal{G}) = \mathbf{P}(A \parallel \mathcal{G})$, the conditional probability function of A with respect to sigma-field \mathcal{G} .

PROPOSITION 2.1. Let $\mathcal{G}_{\mathcal{A}}$ be generated by a countable partition of \mathcal{X} according to (8). If $\mathcal{G}_{\mathcal{I}} \subset \mathcal{G}_{\mathcal{A}} \subset \mathcal{G} \subset \mathcal{F}$, the following follows:

- (1) If $A \in \mathcal{G}_{\mathcal{I}}$, then $\mathbf{P}_0(A \parallel \mathcal{G}_{\mathcal{I}}) = \mathbf{P}(A \parallel \mathcal{G}_{\mathcal{A}}) = \mathbb{1}_A$, a.s.
- (2) If $A \in \mathcal{G}_{\mathcal{A}} \setminus \mathcal{G}_{\mathcal{I}}$, then $\mathbb{1}_A = \mathbf{P}(A \parallel \mathcal{G}_{\mathcal{A}}) = \mathbf{P}_0(A \parallel \mathcal{G})$, a.s.
- (3) The function $\mathbf{E}_{\mathbf{P}}(g \parallel \mathcal{G}_{\mathcal{A}})$ is piecewise constant over all sets A_i in (8) that generate $\mathcal{G}_{\mathcal{A}}$. If additionally $\mathbf{P}(A_i) \neq \mathbf{P}_0(A_i)$ and $\mathbf{E}_{\mathbf{P}}(g \parallel \mathcal{G}_{\mathcal{A}})$ is nonzero on A_i , then $\int_{A_i} \mathbf{E}_{\mathbf{P}}(g \parallel \mathcal{G}_{\mathcal{A}}) d\mathbf{P} \neq \int_{A_i} \mathbf{E}_{\mathbf{P}_0}(g \parallel \mathcal{G}_{\mathcal{A}}) d\mathbf{P}_0$.
- (4) $\mathbf{P}(\mathbf{T}) = \mathbf{E}_{\mathbf{P}}[\mathbf{E}_{\mathbf{P}_0}(f_p \parallel \mathcal{G}_{\mathcal{A}})]$.
- (5) If $\mathcal{G}_{\mathcal{I}} = \{\emptyset, \mathcal{X}\}$, $\mathbf{P}_0(\mathbf{T}) = \mathbf{E}_{\mathbf{P}_0}(f_p \parallel \mathcal{G}_{\mathcal{I}})$ a.s.

The first part of Fact (1) ($\mathbf{P}_0(A \parallel \mathcal{G}_{\mathcal{I}}) = \mathbb{1}_A$) implies that the ignorant agent \mathcal{I} , within his lower discernment $\mathcal{G}_{\mathcal{I}}$, has the *potential* of knowing with certainty whether an event $A \in \mathcal{G}_{\mathcal{I}}$ happened (i.e. $x_0 \in A$) or not, by appropriate choice of \mathbf{P}_0 . Consequently, Fact (1) implies that if \mathcal{I} has the potential to know A with certainty, so does agent \mathcal{A} with his additional discernment. Fact (2) says that had the ignorant agent \mathcal{I} at least the same discernment as \mathcal{A} , he would have the *potential* to know with certainty whether any event $A \in \mathcal{G}_{\mathcal{A}}$ within \mathcal{A} 's discernment happened or not. Fact (3) says that, despite the LHS and RHS of (13) being equal with probability 1, their integrals with respect to \mathbf{P} and \mathbf{P}_0 can be different. Together with Fact (2), it says that the conditional probability function of A can have different integrals under \mathbf{P} than under \mathbf{P}_0 . Facts (4) and (5) are applications of the tower property. Example 7 of Section 4 shows that discernment according to Definition 2.1 cannot be extended to a σ -field $\mathcal{G}_{\mathcal{A}}$ that is not generated from a countable partition (8).

2.2. Learning and knowledge. Let us now formulate learning and knowledge in terms of active information. Learning of proposition p is defined as follows.

DEFINITION 2.2. There is **learning** about p , compared to an ignorant person, if

$$(14) \quad \begin{cases} 0 < I^+(\mathbf{T}) \text{ and } p \text{ is true in the true world } x_0, \\ 0 > I^+(\mathbf{T}) \text{ and } p \text{ is false in the true world } x_0. \end{cases}$$

There is **full learning** about p (regardless of the beliefs of the ignorant person) if $\mathbf{P}(\mathbf{T}) = 1$ when p is true in the true world x_0 , or if $\mathbf{P}(\mathbf{T}) = 0$ when p is false in the true world x_0 .

REMARK 1. In words, agent \mathcal{A} has learned about proposition p , compared to an ignorant agent \mathcal{I} , either when p is true and \mathcal{A} 's posterior belief about p is higher than the prior or when p is false and \mathcal{A} 's posterior belief about p is smaller than the prior.

The agent \mathcal{A} has fully learned p (regardless of the beliefs of the ignorant person) if the posterior belief \mathbf{P} about p is 1 when p is true or 0 when p is false.

An agent in a maximum state of ignorance is represented by a maximum entropy (maxent) distribution \mathbf{P}_0 over \mathcal{X} . However, the notion of learning a proposition is limited, as it does not necessarily entail a particular belief about the true world. Therefore, it does not satisfy the conditions of a *justified* true belief, which requires having a belief *for the right reasons*. Knowledge acquisition is defined to cover this gap.

DEFINITION 2.3. Agent \mathcal{A} has acquired **knowledge** about p , compared to an ignorant person \mathcal{I} , if the following three conditions hold:

K1 The criteria of (14) in Definition 2.2 are satisfied.

K2 $x_0 \in \text{supp}(\mathbf{P})$, the support of \mathbf{P} .

K3 For all $\epsilon > 0$, the closed ball $B_\epsilon[x_0] := \{x \in \mathcal{X} : d(x, x_0) \leq \epsilon\}$ is such that $I^+(B_\epsilon[x_0]) \geq 0$, with strict inequality for some $\epsilon > 0$, where d is a metric over \mathcal{X} .

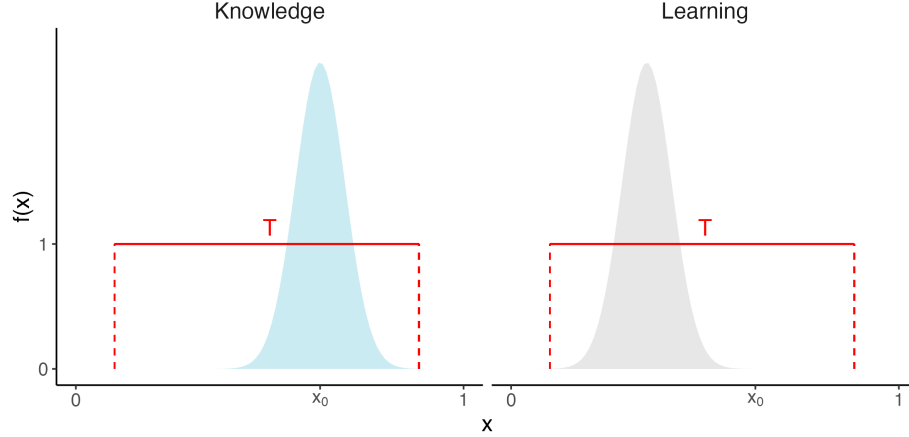


FIG 1. *Learning versus knowledge: The set of possible worlds is $\mathcal{X} = [0, 1]$, the set of worlds where a given proposition p is true is given by \mathcal{T} , the true world is x_0 , and \mathbf{P}_0 is the uniform measure. Thus $\mathbf{P}_0(\mathcal{T}) = \text{length}(\mathcal{T}) < 1$. The light blue region in the LHS represents the beliefs of an agent \mathcal{A}_1 , whereas the gray region in the RHS represents the beliefs of another agent \mathcal{A}_2 . Since the beliefs of the two agents are fully concentrated in \mathcal{T} , $\mathbf{P}_{\mathcal{A}_1}(\mathcal{T}) = \mathbf{P}_{\mathcal{A}_2}(\mathcal{T}) = 1$. Therefore, the two agents fully learned about proposition p . However, since in the RHS $x_0 \notin \text{supp}(\mathbf{P}_{\mathcal{A}_2})$, agent \mathcal{A}_2 does not acquire knowledge, whereas agent \mathcal{A}_1 does as his beliefs are more concentrated around x_0 than those of the ignorant agent with belief \mathbf{P}_0 . Nonetheless, full knowledge is not possible for \mathcal{A}_1 as $\mathbf{P}_{\mathcal{A}_1}$ is continuous.*

Agent \mathcal{A} has acquired **full knowledge** about p (regardless of the beliefs of the ignorant person) if $\mathbf{P} = \delta_{x_0}$.

Condition **K1** ensures that knowledge acquisition is a more stringent concept than learning, as illustrated by Figure 1. Condition **K2** is mathematically equivalent to saying that \mathcal{A} has a positive belief for every open ball centered at x_0 (i.e., if for all $\epsilon > 0$, $\mathbf{P}(B_\epsilon(x_0)) > 0$, where $B_\epsilon(x_0) := \{x \in \mathcal{X} : d(x, x_0) < \epsilon\}$ is the open ball of radius ϵ centered at x_0), which in turn explains Condition **K3**, that the beliefs of \mathcal{A} are more concentrated around x_0 than those of \mathcal{I} .

REMARK 2. The three requirements **K1-K3** of knowledge acquisition from the first part of Definition 2.3 amount respectively to

- (1) \mathcal{A} has learned about p , compared to the ignorant person.
- (2) The true world x_0 is among the pool of possibilities for \mathcal{A} , according to his posterior beliefs.
- (3) The belief in x_0 under \mathbf{P} is stronger than that under \mathbf{P}_0 .

The following result gives sufficient conditions for not having full learning (i. and iii.) and full knowledge acquisition (v.), and sufficient conditions for obtaining full learning (ii. and iv.) and full knowledge acquisition (vi.). In all cases, this is *regardless* of the data \mathcal{D} that agent \mathcal{A} receives. In particular, conditions i. and iii. for not having full learning are such that the truth function f_p of proposition p in (2) is not $\mathcal{G}_{\mathcal{A}}$ -measurable.

THEOREM 2.4. *For the topological space $(\mathcal{X}, \mathcal{O})$, consider the measurable space $(\mathcal{X}, \mathcal{F})$, where $\mathcal{F} = \sigma(\mathcal{O})$. Let \mathbf{P}_0 be a probability measure on $(\mathcal{X}, \mathcal{F})$ and define another probability measure \mathbf{P} on $(\mathcal{X}, \mathcal{F})$ as in (5), where \mathbf{P}_0 and \mathbf{P} represent beliefs about the true world $x_0 \in \mathcal{X}$ of two agents \mathcal{I} and \mathcal{A} respectively. Assume that \mathbf{P}_0 and \mathbf{P} are measurable*

with respect to σ -fields \mathcal{G}_T and \mathcal{G}_A on \mathcal{X} , with $\mathcal{G}_T \subsetneq \mathcal{G}_A \subset \mathcal{F}$. Assume further that $\mathcal{G}_A = \sigma(\mathcal{P})$ is generated from a countable partition \mathcal{P} , according to (8)-(9), such that $\mathbf{P}_0(A_i) > 0$ for all $A_i \in \mathcal{P}$ and none of the $A_i \in \mathcal{P}$ is \mathcal{G}_T -measurable. Let p be a proposition that is true in a set of worlds $T \in \mathcal{F}$, defined in (3). Then

- i. If for all $A \in \mathcal{P}$, it holds that $A \not\subset T$ and $\mathbf{P}_0(A \setminus T) > 0$, then $\mathbf{P}(T) < 1$. In particular, if p is true in the true world x_0 , this implies that full learning of p is not possible.
- ii. Suppose i. fails in the sense that there is an $A \in \mathcal{P}$ such that $A \subset T$. Then we can choose x_0 so that p is true in x_0 , and \mathbf{P} according to (10), so that there is full learning of p , i.e. $\mathbf{P}(T) = 1$.
- iii. If for all $A \in \mathcal{P}$, it holds that $T \cap A \neq \emptyset$ and $\mathbf{P}_0(T \cap A) > 0$, then $\mathbf{P}(T) > 0$. In particular, if p is false in the true world x_0 , this implies that full learning of p is not possible.
- iv. Suppose iii. fails in the sense that there is $A \in \mathcal{P}$ such that $A \cap T = \emptyset$. Then we can choose x_0 such that p is false in x_0 , and \mathbf{P} according to (10), so that there is full learning of p , i.e. $\mathbf{P}(T) = 0$.
- v. If there is $A \in \mathcal{P}$ such that $\{x_0\} \subsetneq A$ and $\mathbf{P}_0(A \setminus \{x_0\}) > 0$, then $\mathbf{P}(\{x_0\}) < 1$ and full knowledge acquisition of not possible.
- vi. If $\{x_0\} \in \mathcal{P}$, then it is possible to choose \mathbf{P} according to (10) such that $\mathbf{P}(x_0) = 1$.

REMARK 3. The conditions imposed in Theorem 2.4 are, in general, easy to obtain, and the result is true with great generality. Note in particular the following:

- Note that $\mathcal{G}_T = \sigma(\mathcal{P}_0)$ is generated from a partition \mathcal{P}_0 coarser than \mathcal{P} , with $\mathbf{P}_0(A) > 0$ for all $A \in \mathcal{P}_0$. Since \mathbf{P}_0 is measurable with respect to \mathcal{G}_T , the conditional distribution of \mathbf{P}_0 is uniform over all $A \in \mathcal{P}_0$. This implies that the conditional distribution of \mathbf{P}_0 is uniform over all sets $A \in \mathcal{P}$ of the finer partition as well.
- Suppose $\mathcal{X} = \mathbb{R}$, $A = [a, b] \in \mathcal{P}$, $T = (a, b)$, and that \mathbf{P}_0 is absolutely continuous with respect to the Lebesgue measure on \mathbb{R} . Then, full learning can be obtained in Theorem 2.4.i. even if $T \subset A$. Thus the requirement that $\mathbf{P}_0(A \setminus T) > 0$ for all $A \in \mathcal{P}$.

3. Maximum entropy and Gibbs posterior distributions.

3.1. *Default choice of posterior.* We will construct the posterior distribution \mathbf{P} in (5) from the prior distribution \mathbf{P}_0 , using $\mathbf{f} = (f_1, \dots, f_n)$ a set of n feature functions $f_i : \mathcal{X} \rightarrow \mathbb{R}$, $i = 1, \dots, n$, with $f_i(X)$ the value of feature i for some randomly generated $X \in \mathcal{X}$. The probability measure \mathbf{P} is generated from \mathbf{P}_0 in such a way that outcomes in regions of \mathcal{X} where f_i is large are either more or less likely under \mathbf{P} compared to \mathbf{P}_0 , given that the other $n - 1$ features do not change. Let

$$(15) \quad \mu_i(\mathbf{P}) = \mathbf{E}_{\mathbf{P}} f_i(X) = \mu_i$$

represent the expected value of feature $i = 1, \dots, n$ under \mathbf{P} , and put $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$. Define \mathcal{Q} as the set of probability measures on \mathcal{X} . It is assumed that

$$(16) \quad \mathbf{P} = \arg \inf_{\mathbf{Q} \in \mathcal{Q}(\boldsymbol{\mu})} D(\mathbf{Q} \parallel \mathbf{P}_0)$$

minimizes the Kullback-Leibler divergence $D(\mathbf{Q} \parallel \mathbf{P}_0) = \mathbf{E}_{\mathbf{Q}} \log[\mathbf{Q}(X)/\mathbf{P}_0(X)]$ (or equivalently maximizes the entropy relative \mathbf{P}_0) among all probability distributions $\mathbf{Q} \in \mathcal{Q}(\boldsymbol{\mu})$, that is, all probability measures that firstly satisfy $\mathbf{Q} \in \mathcal{Q}$, and secondly

$$(17) \quad \mu_i(\mathbf{Q}) = \mu_i, \quad i = 1, \dots, n.$$

Using Lagrange multipliers, we motivate in Section 7 that the solution to the constrained minimization problem (16)-(17) is the Gibbs distribution

$$(18) \quad \mathbf{P}(x) = \mathbf{Q}_{\boldsymbol{\lambda}}(x) = \frac{\mathbf{P}_0(x) e^{\boldsymbol{\lambda} \cdot \mathbf{f}(x)}}{Z_{\boldsymbol{\lambda}}},$$

with $\lambda = (\lambda_1, \dots, \lambda_n) = \lambda(\mu) \in \mathbb{R}^n$ a vector of dimension n chosen so that (17) holds, whereas

$$(19) \quad Z_\lambda = \int_{\mathcal{X}} \mathbf{P}_0(x) e^{\lambda \cdot \mathbf{f}(x)} dx = \mathbf{E}_{\mathbf{P}_0} e^{\lambda \cdot \mathbf{f}(X)}$$

is a normalizing constant, selected so that \mathbf{Q}_λ is a probability measure. In (19), we interpret $dx = \nu(dx)$ as the Lebesgue measure when \mathcal{X} is a subset of a Euclidean space or as the counting measure when \mathcal{X} is a finite set.

Let $D \in \Delta$ be a data set as in (4) that is informative for the values of the n features. We will assume that the expected features $\mu_i = \mu_i(\mathbf{P}) = \mu_i(\mathbf{P}(D))$ in (15) are functions of D . Formally, we may interpret the Gibbs distribution \mathbf{Q}_λ in (18) as a posterior distribution with density

$$(20) \quad \mathbf{P}(x) = \frac{\mathbf{L}(D | x) \mathbf{P}_0(x)}{\mathbf{L}(D)}$$

when the prior distribution is \mathbf{P}_0 and the likelihood is

$$(21) \quad \mathbf{L}(D | x) = e^{\lambda \cdot \mathbf{f}(x)}.$$

Such a connection between Gibbs distributions and Bayesian statistics has been exploited in high-dimensional statistics and statistical physics [1, 44]. When \mathcal{X} is finite or bounded, it is natural to impose a maxent prior \mathbf{P}_0 on \mathcal{X} , equal to the uniform distribution (12). Note also that the formal likelihood in (21) is proportional to a member of an exponential family with parameter $x \in \mathcal{X}$ and sufficient statistic $\lambda = \lambda(D)$ [26]. In particular, x is a natural parameter of this family if $x = \mathbf{f}(x)$, so that the feature extraction does not entail any data reduction. However, (21) is not necessarily an actual likelihood, since

$$\int_{\Delta} \mathbf{L}(\delta | x) d\delta = \int_{\Delta} e^{\lambda(\delta) \cdot \mathbf{f}(x)} d\delta$$

is typically different from 1. The vector λ of the formal likelihood in (21) will be chosen to be consistent with the constraints (17) of the optimization problem (16) that data D in (4) provide.

EXAMPLE 1 (Independent sample). Suppose data $D = (D_1, \dots, D_N)$ is of size N , with components that are observations of independent and identically distributed variables D_1, \dots, D_N . From this it follows that the expected features $\{\mu(D_k)\}_{k=1}^N$ are observations of independent and identically distributed random variables $\mu(D_k)$. If data is unbiased and the second moments of all expected features exist, it follows that $E[\mu(D_k)] = \mathbf{f}(x_0)$ and $\text{Var}[\mu(D_k)] = \Sigma$ where x_0 is the true but unknown value of x , whereas Σ is a covariance matrix of order n . We will also assume that

$$(22) \quad \mu = \mu(D) = \frac{1}{N} \sum_{k=1}^N \mu_k(D)$$

is a sample average of the individually observed feature expectations.

The framework of Example 1 leads to the following asymptotic result for the expected features μ and the posterior distribution \mathbf{P} as N gets large:

THEOREM 3.1. *Suppose the expected features are obtained from an independent sample $D = (D_1, \dots, D_N)$, as defined in (22), and that the other assumptions of Example 1 hold. Then we have weak convergence*

$$(23) \quad \sqrt{N}(\mu(D) - \mathbf{f}(x_0)) \xrightarrow{\mathcal{L}} N(0, \Sigma)$$

as $N \rightarrow \infty$. Let also $\mathbf{P} = \mathbf{P}(\mathbf{D})$ refer to the solution of the optimization problem (16), with $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathbf{D})$. Then

$$(24) \quad \sqrt{N}(\mathbf{P}(\mathbf{D}) - \mathbf{P}_\infty) \xrightarrow{\mathcal{L}} \mathbf{W}$$

as $N \rightarrow \infty$, where \mathbf{P}_∞ is the Gibbs distribution (18) with $\boldsymbol{\mu}(\mathbf{P}_\infty) = \mathbf{f}(x_0)$, whereas \mathbf{W} is a Gaussian signed measure on \mathcal{X} , with $\mathbf{W}(\mathbf{A}) \sim N(0, C(\mathbf{A}, \mathbf{A}))$ and $\text{Cov}(\mathbf{W}(\mathbf{A}), \mathbf{W}(\mathbf{B})) = C(\mathbf{A}, \mathbf{B})$ for all $\mathbf{A}, \mathbf{B} \in \mathcal{F}$, where $C(\mathbf{A}, \mathbf{B})$ is defined in the proof in Section 7.

EXAMPLE 2 (Finite populations). Suppose $\mathcal{X} = \{x_1, \dots, x_d\}$ is a finite set. We can generate \mathcal{X} from a population \mathbf{E} of (a large) size M , which is partitioned into d nonempty subsets $E = \cup_{k=1}^d x_k$, corresponding to a partition $\mathcal{X} = \{x_1, \dots, x_d\}$ of \mathbf{E} . The measurable space $(\mathbf{E}, \sigma(\mathcal{X}))$ consists of all 2^d finite unions of sets x_i , and a distribution \mathbf{Q} on $(\mathbf{E}, \sigma(\mathcal{X}))$ corresponds to probabilities $y_k = \mathbf{Q}(x_k)$ for $k = 1, \dots, d$. It belongs to the $(d-1)$ -dimensional simplex

$$(25) \quad \mathcal{Q} := \left\{ (y_1, \dots, y_d) \in (\mathbb{R}^+)^d : y_1 + \dots + y_d = 1 \right\},$$

where \mathbb{R}^+ is the set of nonnegative real numbers. The distribution

$$(26) \quad \mathbf{P} = \{p_1, \dots, p_d\} \in \mathcal{Q}$$

that is in maxent relative to $\mathbf{P}_0 = \{p_{01}, \dots, p_{0d}\}$ is the Gibbs distribution

$$(27) \quad p_k = \frac{p_{0k} e^{\boldsymbol{\lambda} \cdot \mathbf{f}(x_k)}}{Z_{\boldsymbol{\lambda}}}, \quad k = 1, \dots, d,$$

with $\boldsymbol{\lambda}$ chosen to be consistent with side constraints (17). Since \mathcal{X} is finite, without further background information, and in accordance with (12), we impose a uniform and maxent prior \mathbf{P}_0 , that is, $p_{0k} = 1/d$ so that p_k in (27) becomes

$$(28) \quad p_k = \frac{e^{\boldsymbol{\lambda} \cdot \mathbf{f}(x_k)}}{\sum_{l=1}^d e^{\boldsymbol{\lambda} \cdot \mathbf{f}(x_l)}}, \quad k = 1, \dots, d.$$

Since the simplex \mathbf{P} in (26) is $(d-1)$ -dimensional, the number of features of the Gibbs distribution (27) must satisfy $1 \leq n \leq d-1$ in order to avoid overparametrization.

3.2. Biased choice of posterior. The likelihood (21) can be motivated in terms of giving a posterior that is in maxent relative to \mathbf{P}_0 . For these reasons, we will regard (21) and (12) as the default choices of likelihood and prior. Let

$$(29) \quad \tilde{\mathbf{P}}(x) = \frac{\tilde{\mathbf{L}}(\mathbf{D} | x) \tilde{\mathbf{P}}_0(x)}{\tilde{\mathbf{L}}(\mathbf{D})}$$

be a posterior distribution, whose likelihood $\tilde{\mathbf{L}}$ and prior $\tilde{\mathbf{P}}_0$ are possibly different from those in (21) and (12), respectively. Following [32–34] to measure bias in algorithms, and [17, 23, 45] to measure the bias of prevalence estimators of COVID-19, we use active information to measure bias:

$$(30) \quad \text{Bias}(\mathbf{A}; \mathbf{P}, \tilde{\mathbf{P}}) = I^+(\mathbf{A}; \mathbf{P}, \tilde{\mathbf{P}}) = I^+(\mathbf{A}; \mathbf{P}_0, \tilde{\mathbf{P}}) - I^+(\mathbf{A}; \mathbf{P}_0, \mathbf{P}) = \log \frac{\tilde{\mathbf{P}}(\mathbf{A})}{\mathbf{P}(\mathbf{A})}$$

towards $\mathbf{A} \in \mathcal{X}$ by considering $\tilde{\mathbf{P}}$ instead of \mathbf{P} . In particular, when only the likelihood is misspecified as

$$(31) \quad \tilde{\mathbf{L}}(\mathbf{D} | x) = e^{\tilde{\boldsymbol{\lambda}} \cdot \mathbf{f}(x)}$$

for some $\tilde{\lambda} \neq \lambda$, it follows that

$$(32) \quad \text{Bias}(\mathcal{A}; \lambda, \tilde{\lambda}) = \log \frac{Z_{\tilde{\lambda}}(\mathcal{A})Z_{\lambda}(\mathcal{X})}{Z_{\lambda}(\mathcal{A})Z_{\tilde{\lambda}}(\mathcal{X})},$$

where $Z_{\lambda}(\mathcal{A}) = \int_{\mathcal{A}} \mathbf{P}_0(x) e^{\lambda \cdot \mathbf{f}(x)} dx$.

4. Learning and knowledge acquisition for Gibbs distributions. We will now combine the concepts introduced in Sections 2 and 3. That is, we consider learning and knowledge acquisition when agent \mathcal{A} has a Gibbs posterior (18), based on n feature functions f_1, \dots, f_n and data \mathcal{D} in terms of \mathcal{A} 's expected beliefs $\mu_i = \mu_i(\mathcal{D})$ about the values of the n features. Since \mathcal{A} forms his beliefs about x_0 , based on the largeness/smallness of the feature functions, it is reasonable to define

$$(33) \quad \mathcal{G}_{\mathcal{A}} = \sigma(f_1, \dots, f_n)$$

as the smallest σ -field that makes all feature functions f_1, \dots, f_n measurable. Indeed, we deduce from (12) and (21) that \mathcal{A} 's likelihood as well as his posterior density $\mathbf{P}(x)$ in (18) are both measurable with respect to (33). When the feature functions f_i are binary indicator functions of subsets of \mathcal{X} , then (33) reduces to a finite partition version $\mathcal{G}_{\mathcal{A}} = \sigma(A_1, \dots, A_l)$ of (8), where $n \leq l \leq 2^n$ is the number of non-empty intersections the sets $\{f_i^{-1}(0), f_i^{-1}(1); i = 1, \dots, n\}$. In particular, $l = n$ when the sets $f_i^{-1}(1)$ form a finite partition of \mathcal{X} .

It follows from (18) that each feature i contributes to increase/decrease \mathcal{A} 's beliefs about x_0 in regions where $\lambda_i f_i(x)$ is large/small. This has an impact on learning about a proposition p that is true whenever the value of feature i is at least a constant value f_0 . This corresponds to a truth function $f_p(x) = \mathbb{1}_{\mathcal{T}}(x)$, with

$$(34) \quad \mathcal{T} = \{x \in \mathcal{X} : f_i(x) \geq f_0\}$$

the set of worlds in which p is true. Proposition 4.1 provides details about learning p .

PROPOSITION 4.1. *Consider a proposition p which is true in the set of worlds (34). Assume further that*

$$(35) \quad \min_{x \in \mathcal{X}} f_i(x) \leq f_0 \leq \max_{x \in \mathcal{X}} f_i(x),$$

with at least one of the two inequalities being strict. Then $\mathbf{P}(\mathcal{T}) = \mathbf{Q}_{\lambda}(\mathcal{T})$ is a strictly increasing function of λ_i , with

$$(36) \quad \begin{aligned} \lim_{\lambda_i \rightarrow -\infty} \mathbf{Q}_{\lambda}(\mathcal{T}) &= 0, \\ \lim_{\lambda_i \rightarrow \infty} \mathbf{Q}_{\lambda}(\mathcal{T}) &= 1 \end{aligned}$$

when the other $n - 1$ components of λ are kept fixed. In particular, agent \mathcal{A} learns p (in relation to the ignorant person \mathcal{I}), if the two conditions below hold:

- (i) $\lambda_j = 0$ for all $j \in \{1, \dots, n\} \setminus \{i\}$,
- (ii) either $\lambda_i > 0$ and $f(x_0) \geq f_0$, or $\lambda_i < 0$ and $f(x_0) < f_0$.

Equation (36) implies that, in principle, it is possible for agent \mathcal{A} to attain full learning about a proposition that is true when one feature exceeds a given threshold. It is enough in this case for \mathcal{A} to have data \mathcal{D} that lead to the appropriate expected beliefs $\mu = \mu(\mathcal{D})$ in the values of the features, and the corresponding sufficient statistic $\lambda = \lambda(\mathcal{D})$ of the likelihood (21), that make $\mathbf{Q}_{\lambda}(\mathcal{T})$ close to 1 (0) when p is true (false). However, as it will be seen in Sections 4.1-4.4, for other types of propositions, neither full learning nor full knowledge acquisition is guaranteed when the number of features is too small.

4.1. *Fundamental limits of knowledge for classification on finite populations.* This section presents examples of LKA for classification over finite populations. Example 3 illustrates with one binary feature that full knowledge might not be possible even if full learning is obtained. Theorem 4.1 generalizes the situation to multiple features, proving that there are fundamental limits for full knowledge acquisition.

EXAMPLE 3 (Finite populations with one binary feature.). Continuing Example 2, recall that \mathbf{E} is a population with M subjects, partitioned into d subsets (or subpopulations)

$$(37) \quad \mathcal{X} = \{x_1, \dots, x_d\}.$$

Assume that the first h subpopulations $\mathbf{N}^c := \{x_1, \dots, x_h\}$ are southern, whereas the remaining $d - h$ subpopulations $\mathbf{N} := \{x_{h+1}, \dots, x_d\}$ are northern. Suppose the only feature function

$$(38) \quad f(x_k) = \mathbb{1}_{\mathbf{N}}(x_k)$$

is an indicator as to whether a subpopulation is northern. Assume that \mathbf{D} provides \mathcal{A} with some information as to whether subject \mathcal{S} resides in a northern subpopulation or not. More precisely, based on data (4), \mathcal{A} believes the probability is $\mu = \mathbf{E}_{\mathbf{P}} f(X)$ that \mathcal{S} lives in a northern subpopulation. The Gibbs distribution (18) simplifies to

$$\mathbf{P}(x_k) = \begin{cases} \frac{1}{h+(d-h)e^\lambda} = \frac{1-\mu}{h}; & k = 1, \dots, h, \\ \frac{e^\lambda}{h+(d-h)e^\lambda} = \frac{\mu}{d-h}; & k = h+1, \dots, d, \end{cases}$$

whereas the σ -field in (33) is

$$\mathcal{G}_{\mathcal{A}} = \{\emptyset, \mathbf{N}, \mathbf{N}^c, \mathcal{X}\}.$$

For a data set of size N , it follows from (23) that $\mu \xrightarrow{p} \mathbb{1}_{\mathbf{N}}(x_0)$ as $N \rightarrow \infty$, where \xrightarrow{p} refers to convergence in probability and $x_0 = x_{k_0}$ is the subpopulation where \mathcal{S} actually lives. Moreover, the limiting posterior distribution \mathbf{P}_∞ of (24), as $N \rightarrow \infty$, is a uniform distribution on \mathbf{N} if $x_0 \in \mathbf{N}$, and a uniform distribution on \mathbf{N}^c if $x_0 \notin \mathbf{N}$.

Consider the proposition

$$p : \mathcal{S} \text{ resides in a northern subpopulation.}$$

The truth function (2) of this proposition equals the feature function (38), i.e. $f_p = f$ which implies that the set of worlds (3) for which p is true is

$$\mathbf{T} = \{x_{h+1}, \dots, x_d\} = \mathbf{N}.$$

Suppose p is true, i.e. $x_0 \in \mathbf{T}$. Whenever $d - h \geq 2$, it follows that knowledge acquisition requires more than learning. Indeed, learning occurs whenever

$$(39) \quad \mathbf{P}(\mathbf{T}) = \mathbf{P}(x_{h+1}) + \dots + \mathbf{P}(x_d) > \frac{d-h}{d} = \mathbf{P}_0(\mathbf{T}),$$

which, by Proposition 4.1 with $n = i = f_0 = 1$, is equivalent to $\lambda > 0$. Define the metric $d(x, y) = \mathbb{1}\{x \neq y\}$ on \mathcal{X} . In particular, full learning is attained when the LHS of (39) equals 1. However, it follows from Condition K3 of Definition 2.3 that, on top of (39), full knowledge acquisition is not possible when $d - h \geq 2$, since

$$(40) \quad \mathbf{P}(x_0) \leq \frac{1}{d-h} < 1.$$

We conclude that knowledge acquisition requires more than learning.

Example 3 motivates Theorem 4.1 below. It gives sufficient and necessary conditions for how large n must be to make it possible for \mathcal{A} to attain full knowledge of any proposition. Therefore, it can be seen as a result of fundamental limits of inference for full knowledge in classification problems.

In what follows, $\lceil x \rceil$ stands for the smallest integer larger or equal to x .

THEOREM 4.1 (Fundamental limits of knowledge). *Consider a finite set (37) with n binary features*

$$(41) \quad f_i(x) = 1_{A_i}(x)$$

that are indicator functions for different subsets A_1, \dots, A_n of \mathcal{X} . If

$$(42) \quad n \geq \lceil \log_2 d \rceil,$$

it is possible to choose the sets A_1, \dots, A_n and constants $\lambda_1, \dots, \lambda_n$ so that full knowledge can be attained about any proposition p . Conversely, if n does not satisfy (42), for any choice of n binary features, it is possible to pick x_0 so that full knowledge acquisition is not possible.

The idea of proof of Theorem 4.1 is related to that of Theorem 2.4: The n binary features (41) give rise to a finite partition of \mathcal{X} . If n is too small, then at least one set of this partition will necessarily have more than one element, making full knowledge acquisition impossible.

4.2. Coordinatewise features. In this section, we consider features that are functions of the coordinates of x . We illustrate with two examples that having enough features is crucial for full learning and knowledge acquisition.

EXAMPLE 4 (One feature per coordinate). Assume that

$$(43) \quad \mathcal{X} = [0, 1]^n = \{x = (x_1, \dots, x_n); 0 \leq x_i \leq 1 \text{ for } i = 1, \dots, n\}$$

is the unit cube in n dimensions, with coordinatewise feature functions

$$f_i(x) = x_i$$

for $i = 1, \dots, n$. Data (4) leads \mathcal{A} to form an expected belief $\mu_i = \mathbf{E}_{\mathbf{P}} f_i(X) = \mathbf{E}_{\mathbf{P}}(X_i)$ about the value of each feature $i = 1, \dots, n$, but not about any dependency structure between these features. The beliefs of \mathcal{A} about x_0 are given by the posterior density

$$(44) \quad \mathbf{P}(x) = \prod_{i=1}^n \mathbf{P}_i(x_i),$$

where

$$(45) \quad \mathbf{P}_i(x_i) = \begin{cases} 1, & \lambda_i = 0, \\ \frac{\lambda_i e^{\lambda_i x_i}}{e^{\lambda_i} - 1}, & \lambda_i \neq 0. \end{cases}$$

We deduce from (44) that \mathcal{A} 's beliefs about the n coordinates of x_0 are independent. In spite of this, it follows from (33) that the discernment σ -field is maximal ($\mathcal{G}_{\mathcal{A}} = \mathcal{F}$). The following theorem proves that neither full learning nor full knowledge is guaranteed for \mathcal{A} .

THEOREM 4.2. *In the setting of Example 4, consider propositions p with*

$$(46) \quad \mathbb{T} = \{x \in [0, 1]^n; f_p(x) = 1\} = \times_{i=1}^n [a_i, b_i],$$

where $0 \leq a_i < b_i \leq 1$ for $i = 1, \dots, n$. For propositions p that satisfy (46) and are true ($x_0 \in \mathbb{T}$), full learning of p is possible for \mathcal{A} if and only if at least one of the two conditions $a_i = 0$ and $b_i = 1$ holds for $i = 1, \dots, n$. Moreover, it is only possible for \mathcal{A} to attain full knowledge about p if, additionally, all coordinates of x_0 are either 0 or 1.

EXAMPLE 5 (Two features per coordinate). Assume n is even and that $\mathcal{X} = [0, 1]^{n/2}$ is the unit cube in $n/2$ dimensions. For each coordinate x_i , with $i = 1, \dots, n/2$, define one linear and one quadratic feature

$$\begin{aligned} f_{2i-1}(x) &= x_i, \\ f_{2i}(x) &= x_i^2. \end{aligned}$$

Then (44) holds with

$$(47) \quad \mathbf{P}_i(x_i) = \frac{e^{\lambda_{2i-1}x_i + \lambda_{2i}x_i^2}}{\int_0^1 e^{\lambda_{2i-1}t + \lambda_{2i}t^2} dt}.$$

Example 5 motivates the following result:

THEOREM 4.3. *In the setting of Example 5, it is possible, by appropriate choice of λ , to attain full learning and full knowledge of any proposition p such that either a) p is true and x_0 is an inner point of the truth set \mathbb{T} in (3), or b) p is false and x_0 is an inner point of \mathbb{T}^c .*

Theorem 4.3 shows that two features per coordinate make it possible for agent \mathcal{A} to acquire feature data D such that the corresponding choice of $\lambda = \lambda(D)$ leads to full learning and knowledge acquisition of a proposition p . In contrast, Theorem 4.2 reveals that it is typically not possible for \mathcal{A} to acquire full learning and knowledge about p when only one feature per coordinate is available (regardless of the size N of the dataset D). With one feature per coordinate, \mathcal{A} is only able to form beliefs μ_i about the expected value of each coordinate i . In contrast, with two features per coordinate, \mathcal{A} is able to form beliefs about the expected value μ_{2i-1} as well as the variance $\mu_{2i} - \mu_{2i-1}^2$ of the value of each coordinate i . Theorem 4.3 represents the limit when this expected value converges to x_{0i} (component i of x_0), whereas the variance converges to 0. Note in particular that this agrees with the large sample limit of (23), which in the context of Example 5 reads $\mu_{2i-1} \xrightarrow{P} x_{0i}$ and $\mu_{2i} \xrightarrow{P} x_{0i}^2$ as $N \rightarrow \infty$ for $i = 1, \dots, n/2$. Note also that the limiting posterior distribution \mathbf{P}_∞ in (24) is the point mass δ_{x_0} in Example 5, but a non-degenerate distribution in Example 4.

4.3. *Piecewise constant posterior:* In this section, we present two examples with features that lead to piecewise constant posterior densities \mathbf{P} . As will be seen, for this class of features, full knowledge acquisition is never possible, although full learning is sometimes possible.

EXAMPLE 6 (Piecewise constant posterior in one dimension.). Suppose $\mathcal{X} = (0, 1]$ is the unit interval, which is divided into n equally large and disjoint sets $A_i = ((i-1)/n, i/n]$ for $i = 1, \dots, n$. The feature functions

$$(48) \quad f_i(x) = \mathbb{1}_{A_i}(x)$$

are indicator functions for these intervals. Data D in (4) provides \mathcal{A} with information about the expected features $\mu_i = \mathbf{E}_{\mathbf{P}} f_i(X) = \mathbf{P}(A_i)$ for $i = 1, \dots, n$. Assume also that the ignorant agent \mathcal{I} has a uniform density $\mathbf{P}_0(x) = 1$ on \mathcal{X} , according to (12). From this, it follows that the posterior density (18) of \mathcal{A} is piecewise constant

$$(49) \quad \mathbf{P}(x) = \sum_{i=1}^n p_i \mathbb{1}_{A_i}(x)$$

over each A_i , as in (10), with values

$$(50) \quad p_i = n\mu_i = \frac{ne^{\lambda_i}}{e^{\lambda_1} + \dots + e^{\lambda_n}} \propto e^{\lambda_i}.$$

Note that the feature functions are linearly dependent:

$$(51) \quad \sum_{i=1}^n f_i(x) = 1.$$

For this reason, one of them is redundant. Nonetheless, it is still convenient to have n (rather than $n - 1$) feature functions because of symmetry. The linear dependency (51) implies, however, that λ does not uniquely characterize \mathbf{P} since we may add the same constant to all λ_i without changing \mathbf{P} . Without loss of generality, we can therefore assume that λ is chosen so that the last proportionality of (50) is an equality, which implies that $n = e^{\lambda_1} + \dots + e^{\lambda_n}$. We conclude from (33) and (48) that

$$\mathcal{G}_{\mathcal{A}} = \sigma(A_1, \dots, A_n)$$

is the set of all 2^n finite unions of sets A_i (this corresponds to a finite partition of \mathcal{X} of size n in (9), in order to generate $\mathcal{G}_{\mathcal{A}}$). Hence, $1/n$ is the maximal resolution by which \mathcal{A} is able to discern between different possible worlds. The proposition

$$p : x_0 \text{ belongs to } A_i$$

has truth function $f_p = f_i$, and the set of worlds for which p is true is $T = A_i$. Suppose p is true. It is possible then for \mathcal{A} to fully learn p . This happens when $\mu_i = 1$, or equivalently $p_i = n$, corresponding to the large ($N \rightarrow \infty$) sample limit of (23)-(24), with \mathbf{P}_{∞} the uniform distribution on A_i . But since \mathcal{A} only knows $x_0 \in A_i$, he still has not acquired full knowledge about p . Indeed, in spite of the fact that $\mu_i = 1$, it follows from (49), that for any $\varepsilon < 1/(2n)$, for $B = B(x_0, \varepsilon)$,

$$(52) \quad \mathbf{P}(B) = 1 - \mathbf{P}(B^c) = 1 - n|A_i \setminus B| \leq 1 - n\varepsilon.$$

Suppose $n = 10$, with x_0 the observed value of a uniformly distributed random variable $X \in \mathcal{X}$. In this case, A_i is the event that the first decimal of X is $i - 1$, and data D provide agent \mathcal{A} with information about the first decimal of X . Consider the proposition

$$p' : \text{The second decimal of } X \text{ is } 5,$$

with T' the set of worlds for which p' is true. It is clear that

$$\mathbf{P}(T') = \mathbf{P}_0(T') = 0.1,$$

regardless of the choice of \mathbf{P} in (49). For this reason, \mathcal{A} does not learn anything about p' (the second decimal of X), no matter how accurate information he receives about the first decimal of X . This is an illustration of Theorem 2.4, where it is not only impossible for \mathcal{A} to learn p' fully, but it is not even possible for \mathcal{A} to learn anything at all about p' . In order for \mathcal{A} to learn about p' , he needs to acquire data about the second decimal of X , corresponding to $n = 100$ features. This makes it possible for \mathcal{A} to fully learn p' , although he still does not acquire full knowledge about p' (cf. (52)).

Next, we generalize Example 6 by considering an r -dimensional piecewise constant posterior that is obtained from a recursively partitioned binary tree. Its significance arises from the fact that this is the structure used in the construction of classification and regression trees [4, 36].

THEOREM 4.4. *Let $\mathcal{X} = [0, 1]^r$ and $\mathcal{P} = \{A_1, \dots, A_n\}$ be a finite partition of \mathcal{X} that is obtained as a recursively partitioned binary tree, so that all A_i are rectangles with sides parallel to the coordinate axes. Then, full knowledge is only attained if the number of features n goes to infinity.*

The details of the construction of the recursively partitioned binary tree and the corresponding posterior distribution \mathbf{P} are given in the proof in Section 7.

4.4. *Limits of discernment.* In Example 7 below, we present a σ -field that turns out to be inappropriate for representing the discernment of agent \mathcal{A} . Instead, we will approximate this σ -field with a smaller one that is a mixture of different piecewise constant features, as in Example 6. This selection of features represents information loss, but we introduce it to make it possible for \mathcal{A} to form beliefs.

EXAMPLE 7 (Countable and cocountable sets.). Billingsley presents the following example [2, Example 33.11]: Consider the probability space $(\mathcal{X}, \mathcal{F}, \mathbf{Q})$, where $\mathcal{X} = [0, 1]$, \mathcal{F} is the Borel σ -field on $[0, 1]$, and \mathbf{Q} a continuous probability measure. Consider an agent \mathcal{A} whose discernment $\mathcal{G}_{\mathcal{A}}$ is given by the countable-cocountable subsets of $[0, 1]$ (i.e., $B \in \mathcal{G}_{\mathcal{A}}$ if and only if either B is countable or B^c is countable). Then, for all $A \in \mathcal{F}$,

$$(53) \quad \mathbf{Q}(A) = \mathbf{Q}(A \parallel \mathcal{G}_{\mathcal{A}}),$$

a.s., since $\mathbf{Q}(A)$ is $\mathcal{G}_{\mathcal{A}}$ -measurable, integrable, and it satisfies the functional equation

$$\int_B \mathbf{Q}(A) \mathbf{Q}(dx) = \mathbf{Q}(A) \mathbf{Q}(B) = \mathbf{Q}(A \cap B) = \int_B \mathbf{Q}(A \parallel \mathcal{G}_{\mathcal{A}})(x) \mathbf{Q}(dx)$$

for all $B \in \mathcal{G}_{\mathcal{A}}$. This follows since both sides are either 0 or $\mathbf{Q}(A)$, depending on whether B or B^c is countable. However, every singleton of $[0, 1]$ is $\mathcal{G}_{\mathcal{A}}$ -measurable. Therefore, seeing $\mathcal{G}_{\mathcal{A}}$ as discernment, since A is a union of singletons we would intuitively expect that

$$(54) \quad \mathbb{1}_A = \mathbf{Q}(A \parallel \mathcal{G}_{\mathcal{A}}).$$

However, this intuition goes wrong whenever $\mathbf{Q}(A) > 0$, so that the union is uncountable. Indeed, taking (54) together with (53), we obtain $\mathbf{Q}(A) = \mathbf{Q}(A \parallel \mathcal{G}_{\mathcal{A}}) = \mathbb{1}_A$, a contradiction for all A such that $\mathbf{Q}(A) > 0$. This example shows that the condition of having $\mathcal{G}_{\mathcal{A}}$ generated from a countable partition (8) of \mathcal{X} cannot be removed from Definition 2.1.

We will simplify Billingsley's problem and construct a σ -field $\mathcal{G}_{\mathcal{A}}$ consistent with Definition 2.1. This is contained in the following result, which is proved in Section 7:

PROPOSITION 4.2.

1. Let $A = \{x_1, x_2, \dots\} \subset [0, 1]$ be a fixed countable set, and define

$$(55) \quad \mathcal{G}_{\mathcal{A}} = \sigma([0, 1] \setminus A, x_1, x_2, \dots)$$

as the σ -field generated by the complement of A and the elements of A (or equivalently, the collection of sets B such that either B or B^c is a subset of A). Even though it is not possible to express the posterior as a Gibbs distribution, it is sometimes possible to fully learn and acquire full knowledge about a proposition p with the truth set T (cf. (3)). Full learning is possible if either p is true and $A \cap T \neq \emptyset$ or if p is false and $A \cap T^c \neq \emptyset$. Full knowledge can be attained if additionally p is true and $x_0 \in A \cap T$, or if p is false and $x_0 \in A \cap T^c$.

2. Let

$$(56) \quad \tilde{\mathcal{G}}_{\mathcal{A}} = \sigma([0, 1] \setminus A_n, x_1, x_2, \dots, x_n)$$

be constructed from the finite set $A_n = \{x_1, \dots, x_n\}$. Then, it is possible to approximate the posterior with a Gibbs distribution of n features. Full learning and full knowledge acquisition are possible under the same conditions as in Part 1, with A_n in place of A .

5. Empirical side constraints and secondary learning. In this section, we analyze secondary learning, whereby an agent $\tilde{\mathcal{A}}$ learns about the learning of another agent \mathcal{A} . Whereas agent \mathcal{A} has primary data \mathbf{D} from the n selected features, agent $\tilde{\mathcal{A}}$ has secondary data $\tilde{\mathbf{D}}$ about \mathcal{A} 's learning. In particular, this implies that the interpretation of λ in the Gibbs distribution (18) differs between \mathcal{A} and $\tilde{\mathcal{A}}$. For agent \mathcal{A} , $\lambda = \lambda(\mathbf{D})$ is a sufficient statistic for doing inference about the parameter x , based on the feature data \mathbf{D} that he receives. On the other hand, for agent $\tilde{\mathcal{A}}$, λ is a parameter of \mathcal{A} 's posterior beliefs that needs to be estimated.

As a preparation, we will first, in Section 5.1, introduce optimization (maximum likelihood estimation of λ) under empirical (secondary type of learning) side constraints.

5.1. Optimization under empirical side constraints. A variant of the optimization problem (16)-(17) is to assume that a sample

$$(57) \quad \tilde{\mathbf{D}} = \{x_1, \dots, x_m\}$$

of size m is available from \mathcal{X} , with empirical distribution $\pi = \sum_{j=1}^m \delta_{x_j}/m$, where δ_x refers to a point mass at x . Replace (17) with empirical constraints

$$(58) \quad \mu_i(\mathbf{Q}) = \mu_i(\pi) = \frac{1}{m} \sum_{j=1}^m \Delta_{ij}, \quad i = 1, \dots, n,$$

where $\Delta_{ij} = f_i(x_j)$ is the j -th observed value of feature i . It has been shown in [35] that the solution to the maximization problem (16) is given by $\tilde{\mathbf{P}}(x) = \mathbf{Q}_{\hat{\lambda}}(x)$, where

$$(59) \quad \begin{aligned} \hat{\lambda} &= \arg \max_{\lambda \in \mathbb{R}^n} \prod_{j=1}^m \mathbf{Q}_{\lambda}(x_j) \\ &= \arg \max_{\lambda \in \mathbb{R}^n} \prod_{x \in \mathcal{X}} \mathbf{Q}_{\lambda}(x)^{m\pi(x)} \\ &= \arg \max_{\lambda \in \mathbb{R}^n} \sum_{x \in \mathcal{X}} \pi(x) \log \mathbf{Q}_{\lambda}(x), \\ &= \arg \max_{\lambda \in \mathbb{R}^n} \mathbf{E}_{\pi}(\log \mathbf{Q}_{\lambda}(X)) \\ &= \arg \min_{\lambda \in \mathbb{R}^n} D(\pi \parallel \mathbf{Q}_{\lambda}) \end{aligned}$$

is the maximum likelihood estimator of λ , when $\tilde{\mathbf{D}}$ is viewed as a sample of independent and identically distributed observations from the Gibbs distribution (18). From the third step of (59) we find that $\mathbf{Q}_{\hat{\lambda}}$ is the Gibbs distribution that maximizes the cross entropy between π and \mathbf{Q}_{λ} . This is equivalent to saying that $\mathbf{Q}_{\hat{\lambda}}$ minimizes the expected log loss $\mathbf{E}_{\pi}[-\log \mathbf{Q}_{\lambda}(X)]$ among all Gibbs distributions. It has further been noted (see, e.g., [3, 5, 20]) that the following are convex optimization programs equivalent to those in (59):

$$(60) \quad \begin{aligned} \hat{\lambda} &= \arg \max_{\lambda \in \mathbb{R}^n} \mathbf{E}_{\pi} \left(\log \frac{\mathbf{Q}_{\lambda}(X)}{\mathbf{P}_0(X)} \right) \\ &= \arg \max_{\lambda \in \mathbb{R}^n} [D(\pi \parallel \mathbf{P}_0) - D(\pi \parallel \mathbf{Q}_{\lambda})]. \end{aligned}$$

In particular, from the second step of (60) we deduce that $\hat{\lambda}$ maximizes the expected value $\mathbf{E}_{\pi}[I^+(\{X\}; \mathbf{P}_0, \mathbf{Q}_{\lambda})]$ of an active information measure.

However, recall from Section 1 that $x \in \mathcal{X}$ is the parameter that an agent \mathcal{A} wants to estimate, whereas λ parametrizes the beliefs of \mathcal{A} . Since data $\tilde{\mathbf{D}}$ provide information about λ (and only indirectly about x_0), in the next section we will use it for formalizing the concept of secondary learning.

5.2. *Secondary learning and knowledge acquisition.* Consider an agent $\tilde{\mathcal{A}}$ who does not have access to data D in (4), in order to learn proposition p . However, $\tilde{\mathcal{A}}$ is in contact with agent \mathcal{A} , who has received data D and formed a Gibbs distribution posterior belief $\mathbf{P}(\cdot; \boldsymbol{\lambda})$ about x_0 , according to (18). Agent $\tilde{\mathcal{A}}$ receives a random sample \tilde{D} of size m from \mathcal{A} (as in (57)), drawn from \mathcal{A} 's posterior distribution (18). Based on this, agent $\tilde{\mathcal{A}}$ forms his beliefs about x_0 using either a maximum likelihood approach (Section 5.2.1) or a Bayesian approach (Section 5.2.2) in order to estimate $\boldsymbol{\lambda}$.

5.2.1. *Maximum likelihood plug-in approach.* In this section we assume that $\tilde{\mathcal{A}}$ forms his beliefs about \mathcal{A} 's beliefs about x_0 , from the plug-in posterior distribution

$$(61) \quad \tilde{\mathbf{P}}(x) = \mathbf{P}(x; \hat{\boldsymbol{\lambda}}),$$

where $\hat{\boldsymbol{\lambda}}$ is the maximum likelihood estimator of $\boldsymbol{\lambda}$, defined in (59). It follows from (30) and (32) that agent $\tilde{\mathcal{A}}$ believes that \mathcal{A} has learnt an amount

$$(62) \quad \hat{I}^+(\mathbf{T}) = I^+(\mathbf{T}) + \text{Bias}(\mathbf{T}; \boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$$

about p , where $I^+(\mathbf{T}) = I^+(\mathbf{T}; \mathbf{P}_0, \mathbf{P})$ is the actual amount of learning of \mathcal{A} about p , whereas $\hat{I}(\mathbf{T}) = I^+(\mathbf{T}; \mathbf{P}_0, \tilde{\mathbf{P}})$ is $\tilde{\mathcal{A}}$'s estimate of this quantity. The following proposition gives an asymptotic expansion of $\tilde{\mathcal{A}}$'s expected estimate of \mathcal{A} 's learning:

PROPOSITION 5.1. *Suppose agent $\tilde{\mathcal{A}}$ forms his beliefs about agent \mathcal{A} 's beliefs in x_0 according to (61), based on a secondary learning data set \tilde{D} of size m , drawn randomly from \mathcal{A} 's posterior distribution $\mathbf{P} = \mathbf{Q}_{\boldsymbol{\lambda}}$ in (18). Then asymptotically, $\tilde{\mathcal{A}}$'s expected learning about agent \mathcal{A} 's beliefs in proposition p is*

$$(63) \quad \mathbf{E}[\hat{I}^+(\mathbf{T})] = I^+(\mathbf{T}) + \frac{C}{m} + o(m^{-1})$$

as $m \rightarrow \infty$, where \mathbf{T} is the set of worlds (3) for which p is true, and expectation is taken with respect to random variations in \tilde{D} . Moreover, $C = \text{tr}(\mathbf{J}^{-1}\mathbf{H})/2$, $\mathbf{J} = \mathbf{J}(\boldsymbol{\lambda}) = \mathbf{E}_{\mathbf{Q}_{\boldsymbol{\lambda}}}[\mathbf{f}(X)\mathbf{f}(X)^T]$ is the Fisher information matrix that corresponds to the maximum likelihood estimate (59) of $\boldsymbol{\lambda}$, \mathbf{H} is the Hessian matrix of the function $\boldsymbol{\lambda}' \rightarrow \text{Bias}(\mathbf{T}; \boldsymbol{\lambda}, \boldsymbol{\lambda}')$ at $\boldsymbol{\lambda}' = \boldsymbol{\lambda}$, whereas $o(m^{-1})$ is a remainder term that is small in comparison to m^{-1} as $m \rightarrow \infty$.

5.2.2. *Bayesian approach.* Has $\tilde{\mathcal{A}}$ learned and acquired knowledge about p ? Not necessarily, since $\tilde{\mathcal{A}}$ tries to recapitulate the beliefs of \mathcal{A} about p , based on data \tilde{D} , without having access to original data D in (4) that \mathcal{A} used in order to formulate his beliefs about p . It is safer to say that $\tilde{\mathcal{A}}$ learns and acquires knowledge about how much \mathcal{A} has learned about p . This corresponds to an LKA problem with a true world

$$\tilde{x}_0 = I^+(\mathbf{T}) \in (-\infty, -\log \mathbf{P}_0(\mathbf{T})] =: \tilde{\mathcal{X}}.$$

In order to define this LKA problem properly, in line with Section 2, in this section we take a Bayesian approach about $\boldsymbol{\lambda}$ and treat it as a random parameter with a prior $\mathbf{P}_0(\boldsymbol{\lambda})$ and posterior

$$(64) \quad \tilde{\mathbf{P}}(\boldsymbol{\lambda}) \propto \tilde{\mathbf{L}}(\tilde{D} \mid \boldsymbol{\lambda}) \mathbf{P}_0(\boldsymbol{\lambda}),$$

where $\tilde{\mathbf{L}}(\tilde{D} \mid \boldsymbol{\lambda})$ is the likelihood defined in the first line of (59), used by agent $\tilde{\mathcal{A}}$ in order to make inference about $\boldsymbol{\lambda}$. This gives rise to a modified version

$$(65) \quad \tilde{\mathbf{P}}(x) = \int \mathbf{P}(x; \boldsymbol{\lambda}) \tilde{\mathbf{P}}(d\boldsymbol{\lambda})$$

of (61), that is, a modified version of agent $\tilde{\mathcal{A}}$'s expected beliefs about \mathcal{A} 's beliefs about $x_0 \in \mathcal{X}$. In order to formalize $\tilde{\mathcal{A}}$'s learning about \mathcal{A} 's learning, consider the proposition

\tilde{p} : Agent \mathcal{A} has increased his beliefs that p is true.

This proposition is true if $\tilde{x}_0 = I^+(\mathsf{T}) \in (0, -\log \mathbf{P}_0(\mathsf{T})] := \tilde{\mathsf{T}} \subset \tilde{\mathcal{X}}$. Hence, agent $\tilde{\mathcal{A}}$'s amount of learning about \tilde{p} is given by

$$\tilde{I}^+(\tilde{\mathsf{T}}) = \log \frac{\tilde{\mathbf{P}}(\tilde{\mathsf{T}})}{\mathbf{P}_0(\tilde{\mathsf{T}})},$$

where

$$\begin{aligned} \mathbf{P}_0(\tilde{\mathsf{T}}) &= \int \mathbf{1}\{I^+(\mathsf{T}; \boldsymbol{\lambda}) > 0\} \mathbf{P}_0(d\boldsymbol{\lambda}), \\ \tilde{\mathbf{P}}(\tilde{\mathsf{T}}) &= \int \mathbf{1}\{I^+(\mathsf{T}; \boldsymbol{\lambda}) > 0\} \tilde{\mathbf{P}}(d\boldsymbol{\lambda}), \end{aligned}$$

represent agent $\tilde{\mathcal{A}}$'s beliefs in $\tilde{\mathsf{T}}$ before and after he received data $\tilde{\mathsf{D}}$ respectively. On the right-hand side of the last equation, we made use of the simplified notation $I^+(\mathsf{T}; \boldsymbol{\lambda}) = I^+(\mathsf{T}; \mathbf{P}_0, \mathbf{Q}_{\boldsymbol{\lambda}})$.

In addition, $\tilde{\mathcal{A}}$ also learns and acquires knowledge about how much knowledge \mathcal{A} has acquired about p . This corresponds to an LKA problem with a true world $\tilde{x}_0 = \mathbf{P}(\cdot; \boldsymbol{\lambda}) \in \mathcal{Q} =: \tilde{\mathcal{X}}$, where \mathcal{Q} is the set of distributions on \mathcal{X} . From the posterior distribution (64) of $\boldsymbol{\lambda}$ given data $\tilde{\mathsf{D}}$, it is possible to compute a posterior distribution of the distribution $\mathbf{P}(\cdot; \boldsymbol{\lambda})$ given data $\tilde{\mathsf{D}}$ for agent $\tilde{\mathcal{A}}$. The latter posterior distribution can be used to define various aspects of agent $\tilde{\mathcal{A}}$'s learning and knowledge acquisition about \mathcal{A} 's knowledge acquisition about p .

6. Discussion. In this paper, we used the concept of active information to analyze learning of a proposition and knowledge acquisition of the true world for an agent \mathcal{A} who receives data D in terms of a number of features that are of relevance for learning and knowledge acquisition. This leads to a Gibbs distribution for the posterior distribution of the true world that corresponds to the beliefs of \mathcal{A} . We also introduced the concept of secondary learning for an agent $\tilde{\mathcal{A}}$ who does not have access to original data D but rather receives data $\tilde{\mathsf{D}}$ from \mathcal{A} .

Our work has implications for statistical learning, where an algorithm \mathcal{A} receives data on a number of features of an object x_0 in order to learn and acquire knowledge about various propositions of relevance for the object. We have highlighted potential limitations of such statistical learning algorithms based on feature extraction: When the number of features is too small, this type of primary learning is not always possible, and full knowledge acquisition is not guaranteed.

The results of this article can be extended in various ways. Firstly, one can look at learning and knowledge acquisition dynamically as a function of the size of the data set. This holds for primary data $\mathsf{D} = (\mathsf{D}_1, \dots, \mathsf{D}_N)$ as well as for secondary data $\tilde{\mathsf{D}} = (x_1, \dots, x_m)$. For primary data, under the assumptions of Example 1, it follows from (23) that agent \mathcal{A} 's expected beliefs $\boldsymbol{\mu}$ about the values of the features will converge to $\mathbf{f}(x_0)$, the true values of these features, as $N \rightarrow \infty$. However, if the number of features n is too small (independently of N), full knowledge acquisition will not be possible, even when $N \rightarrow \infty$ (cf. Theorem 3.1). For secondary learning, we expect that asymptotically as $m \rightarrow \infty$, agent $\tilde{\mathcal{A}}$ learns perfectly well about \mathcal{A} 's learning and knowledge acquisition (cf. Proposition 5.1). Because of this, the posterior distribution $\tilde{\mathbf{P}}$ of agent $\tilde{\mathcal{A}}$ will not get concentrated around x_0 as $m \rightarrow \infty$, but rather

converge to the posterior distribution \mathbf{P} of agent \mathcal{A} . For primary and secondary learning, if the data is increased one sample at a time, we can think of the resulting learning process as a Glauber dynamics (Gibbs sampler). This makes it possible to analyze various asymptotic properties of the learning process, such as fast mixing times [27].

Secondly, there are other types of artificial data sets than secondary data $\tilde{\mathbf{D}}$ that can be used for learning and knowledge acquisition. One such example is synthetic primary data \mathbf{D}' . It is possible, for instance, that synthetic primary data is one of the reasons why large language models sometimes produce outputs with high error rates (see, e.g., [6, 28] and references therein). In order to explain this concept, recall that primary data $\mathbf{D} \in \Delta$ is used by agent \mathcal{A} for making inferences about the true world $x_0 \in \mathcal{X}$. This primary data is an observation of a random variable D on Δ , whose distribution is assumed to follow the mixed likelihood $\int \mathbf{L}(\cdot|x_0)\mathbf{P}_0(x)dx$ of agent \mathcal{I} (although the true likelihood, for data generated without bias, is $\mathbf{L}(\cdot|x_0)$). Recall also that secondary data $\tilde{\mathbf{D}} \in \mathcal{X}^m$ is an independent sample of size m , generated by agent \mathcal{A} from the distribution \mathbf{P} on \mathcal{X} that constitutes his beliefs about x_0 . Synthetic primary data, on the other hand, is artificial primary data generated by \mathcal{A} . It can be viewed as an observation of a random variable D' on Δ whose distribution follows the mixed likelihood $\mathbf{L}(\cdot) = \int \mathbf{L}(\cdot|x)\mathbf{P}(x)dx$ of \mathcal{A} . Consequently, \mathbf{D}' and $\tilde{\mathbf{D}}$ are both generated by agent \mathcal{A} , but for the different purposes of producing new (artificial) primary data and informing about the beliefs of \mathcal{A} respectively. In spite of this, synthetic primary data will have similar asymptotic consequences as secondary data. In order to motivate this, assume that synthetic primary data $\mathbf{D}' = (D'_1, \dots, D'_{N'})$ of size N' is available to agent \mathcal{A}' , where D_k are observations of independent and identically distributed random variables D'_k . Then, in the same way as for primary data (22), \mathcal{A}' forms expected features as a sample average

$$\mu(\mathbf{D}') = \frac{1}{N'} \sum_{k=1}^{N'} \mu_k(D'_k).$$

If primary synthetic data are consistent with the beliefs of agent \mathcal{A} , $\mu_k(D'_k)$ are observations of independent and identically distributed random variables $\mu_k(D'_k)$ with $E[\mu_k(D'_k)] = \mu(\mathbf{P})$. Analogously to Theorem 3.1, if we let $N' \rightarrow \infty$ it then follows that $\mu(\mathbf{D}') \xrightarrow{p} \mu(\mathbf{P})$, and consequently $\mathbf{P}' \xrightarrow{p} \mathbf{P}$, since \mathbf{P} is the Gibbs distribution that corresponds to the limiting expected feature $\mu = \mu(\mathbf{P})$. This is to say that the posterior distribution \mathbf{P}' of agent \mathcal{A}' (just as the posterior distribution \mathbf{P}' for agent $\tilde{\mathcal{A}}$) converges to \mathbf{P} rather than a point mass at x_0 . The conclusion is that neither synthetic primary data nor secondary data will generate full knowledge about a proposition as the size of the data set increases, unless agent \mathcal{A} has already acquired full knowledge about this proposition. In particular, for objects that are either rare and/or related to moral, ethical, and religious issues, it seems that synthetic primary learning and secondary learning algorithms are subject to bias since these two types of learning ultimately depend on others learning about the objects rather than the objects themselves. These observations reinforce our claim that statistical learning does not always entail knowledge.

Thirdly, our results have implications for the learning of whether objects are fine-tuned or not. Suppose, for instance, that we have a single feature ($n = 1$) of the Gibbs distribution (18), i.e. a distribution $\mathbf{P}(x) = \mathbf{P}_0(x)e^{\lambda f(x)}/Z_\lambda$ that is an exponentially tilted version of \mathbf{P}_0 . Such a distribution has been used in [12] to model fine-tuning, with $f(x)$ quantifying the amount of tuning of x , and $\mathbf{T} = \{x \in \mathcal{X}; f(x) \geq f_0\}$ a set of outcomes with a large amount of tuning (a special case of (34) for $n = 1$). An exponentially tilted distribution \mathbf{P} with $\lambda > 0$ corresponds to an algorithm that more often generates outcomes with a large amount of tuning compared to chance. In our setting, the set \mathbf{T} is the truth set of a proposition p that object x_0 has a high amount of tuning. Moreover, \mathbf{P}_0 and \mathbf{P} (with $\lambda > 0$) correspond to beliefs of two agents \mathcal{I} and \mathcal{A} , where \mathcal{A} has stronger beliefs than \mathcal{I} that the true structure x_0 is highly tuned.

Fourthly, in our approach to learning and knowledge acquisition, the posterior distribution of agent \mathcal{A} minimizes the Kullback-Leibler divergence to the prior distribution of agent \mathcal{I} among all distributions that satisfy side constraints in terms of expected features. This can be viewed as a method of moments approach, where the moments of the features are used for inference of the posterior distribution. In particular, this approach implies that the likelihood (21) of the posterior distribution is not the actual likelihood of data but rather a solution to an optimization problem. An alternative strategy is to use the true likelihood $\mathbf{L}(D|x)$ of agent \mathcal{A} for data D based on n features in order to define his posterior distribution (20).

7. Mathematical proofs.

PROOF OF PROPOSITION 2.1. For $A \in \mathcal{F}$, let $g := \mathbb{1}_A$. Then (13) implies that

$$(66) \quad \mathbf{P}(A \parallel \mathcal{G}) = \mathbf{P}_0(A \parallel \mathcal{G}),$$

a.s. To prove Fact (1), assume $A \in \mathcal{G}_{\mathcal{I}}$. Then

$$(67) \quad \mathbb{1}_A = \mathbf{P}_0(A \parallel \mathcal{G}_{\mathcal{I}}) = \mathbf{P}_0(A \parallel \mathcal{G}) = \mathbf{P}(A \parallel \mathcal{G}) = \mathbf{P}(A \parallel \mathcal{G}_{\mathcal{A}}),$$

a.s., where the first equality is due to the fact that $\mathbb{1}_A$ is a version of $\mathbf{P}_0(A \parallel \mathcal{G}_{\mathcal{I}})$; the second equality is due to the fact that $A \in \mathcal{G}_{\mathcal{I}} \Rightarrow A \in \mathcal{G}$; the third equality is due to (66); and the last equality is due to the fact that $A \in \mathcal{G}_{\mathcal{A}} \subset \mathcal{G}$, since $A \in \mathcal{G}_{\mathcal{I}}$. Moreover, the first and third equalities in (67) are a.s.

To prove Fact (2), assume $A \in \mathcal{G}_{\mathcal{A}} \setminus \mathcal{G}_{\mathcal{I}}$. Then (66) implies that

$$(68) \quad \mathbb{1}_A = \mathbf{P}(A \parallel \mathcal{G}_{\mathcal{A}}) = \mathbf{P}(A \parallel \mathcal{G}) = \mathbf{P}_0(A \parallel \mathcal{G}).$$

To prove Fact (3), let c_i be the constant value of $\mathbf{E}_{\mathbf{P}}(g \parallel \mathcal{G}_{\mathcal{A}}) = \mathbf{E}_{\mathbf{P}_0}(g \parallel \mathcal{G}_{\mathcal{A}})$ on A_i . Then, since $\mathbf{P}(A_i) \neq \mathbf{P}_0(A_i)$, if $c_i \neq 0$ it follows that

$$(69) \quad \int_{A_i} \mathbf{E}_{\mathbf{P}}(g \parallel \mathcal{G}) d\mathbf{P} = c_i \mathbf{P}(A_i) \neq c_i \mathbf{P}_0(A_i) = \int_{A_i} \mathbf{E}_{\mathbf{P}_0}(g \parallel \mathcal{G}) d\mathbf{P}_0.$$

As for Fact (4),

$$(70) \quad \mathbf{P}(\mathcal{T}) = \mathbf{E}_{\mathbf{P}}(f_p) = \mathbf{E}_{\mathbf{P}}[\mathbf{E}_{\mathbf{P}}(f_p \parallel \mathcal{G}_{\mathcal{A}})] = \mathbf{E}_{\mathbf{P}}[\mathbf{E}_{\mathbf{P}_0}(f_p \parallel \mathcal{G}_{\mathcal{A}})],$$

where the first equality is obtained by definition of f_p , the second is an application of the tower property, and the last one uses (13).

To prove Fact (5) observe that if $\mathcal{G}_{\mathcal{I}} = \{\emptyset, \mathcal{X}\}$, then $\mathbf{E}_{\mathbf{P}_0}(f_p \parallel \mathcal{G}_{\mathcal{I}})$ is constant a.s. The result then follows from a second application

$$\mathbf{P}_0(\mathcal{T}) = \mathbf{E}_{\mathbf{P}_0}(f_p) = \mathbf{E}_{\mathbf{P}_0}[\mathbf{E}_{\mathbf{P}_0}(f_p \parallel \mathcal{G}_{\mathcal{I}})] = \mathbf{E}_{\mathbf{P}_0}(f_p \parallel \mathcal{G}_{\mathcal{I}})$$

of the tower property, with the last identity holding a.s. □

PROOF OF THEOREM 2.4. All parts are proven in order:

i. For each set A_i of the partition \mathcal{P} , define

$$(71) \quad q_i = \mathbf{P}(\mathcal{T} | A_i) = \mathbf{P}_0(\mathcal{T} | A_i) = 1 - \frac{\mathbf{P}_0(A_i \setminus \mathcal{T})}{\mathbf{P}_0(A_i)} < 1,$$

where the last step is a consequence of the assumptions $\mathbf{P}_0(A_i) > 0$ and $\mathbf{P}_0(A_i \setminus \mathcal{T}) > 0$. It follows from the Law of Total Probability that

$$\mathbf{P}(\mathcal{T}) = \sum_i \mathbf{P}(A_i) q_i < \sum_i \mathbf{P}(A_i) = 1,$$

where the inequality was deduced from (71) and the fact that $\mathbf{P}(A_i) > 0$ for at least one i .

- ii. If i_0 is the index for which $A_{i_0} \subset T$, choose $x_0 \in A_{i_0}$ and $P(A_{i_0}) = 1$.
- iii. Note that $T^c = \mathcal{X} \setminus T$ satisfies the conditions of Theorem 2.4.iii.. Hence $P(T^c) < 1$ and $P(T) = 1 - P(T^c) > 0$.
- iv. If i_0 is the index for which $P_0(A_{i_0} \cap T) = 0$, choose $x_0 \in A_{i_0}$ and $P(A_{i_0}) = 1$.
- v. Make $T = \{x_0\}$ in Theorem 2.4.i.. The result follows.
- vi. This is trivial.

□

Motivation that (18) solves the constrained minimization problem (16)-(17).

In order to motivate that the Gibbs distribution (18) is the solution to the minimization problem (16)-(17), we will use Lagrange multipliers. Our goal is to find the distribution $Q \in \mathcal{P}$ that minimizes the loss function

$$(72) \quad \mathcal{L}(Q) = \int_{\mathcal{X}} Q(x) \left[\log \frac{Q(x)}{P_0(x)} - \lambda \cdot f(x) - \xi \right] dx - (\lambda \cdot \mu - \xi),$$

where $\mu = (\mu_1(P), \dots, \mu_n(P))^T$. The minimizer of (72) must satisfy

$$0 = \frac{\partial \mathcal{L}(Q)}{\partial Q(x)} = \log \frac{Q(x)}{P_0(x)} + 1 - \lambda(x) - \xi$$

for all $x \in \mathcal{X}$, with solution

$$(73) \quad Q(x) = P_0(x) \exp(\lambda \cdot f(x) + \xi - 1).$$

The constants λ and ξ are chosen in (73) so that the side constraints (17) and $\int_{\mathcal{X}} Q(x) dx = 1$ are fulfilled, and this is equivalent to (18). □

PROOF OF THEOREM 3.1. Equation (23) follows directly from the Central Limit Theorem. In order to prove (24), write $P(x; \mu)$ for the solution to optimization problem (16), and let $\mu_\infty = f(x_0)$ for the limiting value of $\mu = \mu(D)$ in (23). For each $A \in \mathcal{F}$, we then use the Delta method, that is, a first-order Taylor expansion

$$P(A; \mu) = \sum_{x \in A} P(x; \mu)$$

around the point μ_∞ , according to

$$P(A; \mu) \approx P(A; \mu_\infty) + P'(A; \mu_\infty)(\mu - \mu_\infty)^T,$$

where $P(A; \mu_\infty) = P_\infty(A)$, $P'(A; \mu) = dP(A; \mu)/d\mu$, whereas T refers to vector transposition. Then (24) follows from (23), with

$$C(A, B) = P'(A; \mu_\infty) \Sigma P'(B; \mu_\infty)^T.$$

□

PROOF OF PROPOSITION 4.1. In order to verify that $P(T) = Q_\lambda(T)$ is a strictly increasing function of λ_i , we use the same method of proof as in Proposition 1 of [12]. To this end it is convenient to introduce $\tilde{P} = Q_{\tilde{\lambda}}$, where $\tilde{\lambda} = (\tilde{\lambda}_1, \dots, \tilde{\lambda}_n)$ has components

$$\tilde{\lambda}_j = \begin{cases} \lambda_j; & j \neq i, \\ 0; & j = i. \end{cases}$$

Define

$$(74) \quad \begin{aligned} J(\lambda_i) &= \sum_{x \in T^c} e^{\lambda_i[f(x) - f(x_0)]} \tilde{P}(x), \\ K(\lambda_i) &= \sum_{x \in T} e^{\lambda_i[f(x) - f(x_0)]} \tilde{P}(x), \end{aligned}$$

when \mathcal{X} is countable, and replace the sums in (74) by integrals when \mathcal{X} is continuous. Then

$$\begin{aligned}
 \mathbf{Q}_\lambda(\mathbf{T}) &= \frac{e^{\lambda_i f(x_0)} K(\lambda_i)}{e^{\lambda_i f(x_0)} [J(\lambda_i) + K(\lambda_i)]} \\
 (75) \quad &= \frac{K(\lambda_i)}{J(\lambda_i) + K(\lambda_i)} \\
 &= \frac{1}{\frac{J(\lambda_i)}{K(\lambda_i)} + 1}.
 \end{aligned}$$

Since by assumption f_0 is an inner point of the range of f_i , it follows that $0 < \tilde{\mathbf{P}}(\mathbf{T}) < 1$. From this, we deduce that $J(\lambda_i)$ is a strictly decreasing function of λ_i , and/or $K(\lambda_i)$ is a strictly increasing function of λ_i . This implies that $\mathbf{P}(\mathbf{T}) = \mathbf{Q}_\lambda(\mathbf{T})$ is a strictly increasing function of λ_i . The lower part of (36) follows from the fact that

$$\begin{aligned}
 (76) \quad &\lim_{\lambda_i \rightarrow \infty} J(\lambda_i) = 0, \\
 &\lim_{\lambda_i \rightarrow \infty} K(\lambda_i) = \infty
 \end{aligned}$$

when both inequalities of (35) are strict. If only one of the two inequalities of (35) is strict, then at least one of the two limits of (76) are valid, so that (36) still holds. The upper part of (36) is proved similarly.

The second part of Proposition 4.1 then follows from the definition of learning in Definition 2.2, and the facts that $\mathbf{P} = \mathbf{Q}_\lambda$ and $\tilde{\mathbf{P}} = \mathbf{Q}_{\tilde{\lambda}} = \mathbf{P}_0$ when $\tilde{\lambda} = (0, \dots, 0)$. \square

PROOF OF THEOREM 4.1. The Gibbs distribution (18) takes the form

$$(77) \quad \mathbf{P}(x_k) = \frac{\exp [\sum_{i=1}^n \lambda_i \mathbb{1}_{A_i}(x_k)]}{\sum_{l=1}^d \exp [\sum_{i=1}^n \lambda_i \mathbb{1}_{A_i}(x_l)]}$$

for some constants $\lambda_1, \dots, \lambda_n$ that quantify the impact of each feature on agent \mathcal{A} 's posterior beliefs. In this case, data in (4) provide \mathcal{A} with information about the probability $\mu_i = \mathbf{E}_{\mathbf{P}} f_i(X) = \mathbf{P}(A_i)$ of each set A_i .

We will first show that whenever (42) holds, there are feature functions f_1, \dots, f_n in (41) such that for any $x \in \mathcal{X}$ it is possible to choose the parameter vector $\lambda = \lambda(x)$ of the Gibbs distribution \mathbf{P} in (77), that represents agent \mathcal{A} 's beliefs, so that $\mathbf{P}(\{x\}) = 1$. This will prove the result since, in particular for the true world x_0 , it implies that

$$(78) \quad \mathbf{P}(\{x_0\}) = 1.$$

is equivalent to full knowledge acquisition of \mathcal{A} for any proposition p (see Definition 2.3). With n as in (42) it is possible to write $x_k = (x_{k1}, \dots, x_{kn}) \in \mathcal{X}$ as a binary expansion of the number $k - 1$ for $k = 1, \dots, d$. Then choose the indicator sets of the feature functions (41) as

$$A_i = \{x_k; x_{ki} = 1\}$$

for $i = 1, \dots, n$. Let $x_0 = (x_{01}, \dots, x_{0n})$ be the binary expansion of $x_0 = x_{k_0}$, and let $\lambda > 0$ be a large number. Pick $\lambda = \lambda(x_0) = (\lambda_1, \dots, \lambda_n)$ so that

$$\lambda_i = \begin{cases} \lambda; & \text{if } x_{0i} = 1, \\ -\lambda; & \text{if } x_{0i} = 0. \end{cases}$$

For each $x_k \in \mathcal{X}$ we define the two subsets $I_0(x_k) = \{i; x_{ki} = 0\}$ and $I_1(x_k) = \{i; x_{ki} = 1\}$ of $\{1, \dots, n\}$. It follows from (77) that

$$\mathbf{P}(x_k) = C e^{\lambda n_k}$$

where $n_k = |I_1(x_0) \cap I_1(x_k)| - |I_0(x_0) \setminus I_0(x_k)|$ is an integer and C is a normalizing constant assuring that \mathbf{P} is a probability measure. Since $k \in \{1, \dots, d\} \rightarrow n_k$ is uniquely maximized for $k = k_0$ by $n_{k_0} = |I_1(x_0)|$, equation (78) follows by letting $\lambda \rightarrow \infty$. This completes the proof of the first part of Proposition 4.1.

Assume next that (42) does not hold, so that $n < \log_2 d$ and $2^n < d$. For each binary vector $\mathbf{f} = (f_1, \dots, f_n)$ of length n , define the set

$$(79) \quad \mathbf{B}_{\mathbf{f}} = \{x \in \mathcal{X}; \mathbf{f}(x) = (f_1(x), \dots, f_n(x)) = \mathbf{f}\}.$$

Suppose $d_0 \leq 2^n$ of the 2^n sets in (79) are non-empty. It follows from (77) that agent \mathcal{A} 's posterior density $\mathbf{P}(x)$ is constant on each non-empty set in (79). Since these d_0 non-empty sets form a disjoint decomposition of \mathcal{X} , and $d_0 \leq 2^n < d$, it follows that $|\mathbf{B}_{\mathbf{f}_0}| > 1$ for at least one binary vector \mathbf{f}_0 . If $x_0 \in \mathbf{B}_{\mathbf{f}_0}$ we deduce that $\mathbf{P}(x_0; \boldsymbol{\lambda}) \leq 1/|\mathbf{B}_{\mathbf{f}_0}| \leq 0.5$, regardless of the value of $\boldsymbol{\lambda}$. According to Definition 2.3, full knowledge acquisition is not possible for this particular x_0 . \square

PROOF OF THEOREM 4.2. Recall that \mathcal{A} forms his beliefs according the Gibbs distribution (44)-(45) for some vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$, and that the set \mathbf{T} of worlds for which the proposition p is true is given by (46). Since we assume that p is true ($x_0 \in \mathbf{T}$), it follows from Definition 2.2 that full learning of p is possible if we can find a vector $\boldsymbol{\lambda}$ such that

$$(80) \quad \mathbf{P}(\mathbf{T}; \boldsymbol{\lambda}) \geq 1 - \epsilon$$

for any $\epsilon > 0$. Thus, we need to look more closely at $\mathbf{P}(\mathbf{T}; \boldsymbol{\lambda})$. Equations (44)-(46) imply that

$$(81) \quad \mathbf{P}(\mathbf{T}; \boldsymbol{\lambda}) = \prod_{i=1}^n \int_{a_i}^{b_i} \mathbf{P}_i(x) dx = \prod_{i=1}^n G(a_i, b_i; \lambda_i),$$

where

$$G(a, b, \lambda) = \begin{cases} \frac{e^{\lambda b} - e^{\lambda a}}{e^{\lambda} - 1}; & \text{if } \lambda \neq 0, \\ b - a; & \text{if } \lambda = 0. \end{cases}$$

Maximizing (81) with respect to $\boldsymbol{\lambda}$, it can be seen that

$$(82) \quad \sup_{\boldsymbol{\lambda}} \mathbf{P}(\mathbf{T}; \boldsymbol{\lambda}) = \prod_{i=1}^n \bar{G}(a_i, b_i),$$

where

$$(83) \quad \bar{G}(a, b) = \sup_{\lambda} G(a, b; \lambda) \begin{cases} = 1; & \text{if at least one of } a = 0 \text{ or } b = 1 \text{ holds,} \\ < 1; & \text{otherwise.} \end{cases}$$

We deduce from (82)-(83) that

$$(84) \quad \sup_{\boldsymbol{\lambda}} \mathbf{P}(\mathbf{T}; \boldsymbol{\lambda}) = 1$$

if and only if at least one of the two conditions $a_i = 0$ or $b_i = 1$ holds for $i = 1, \dots, n$. In view of (80), this proves the first (learning) part of the theorem.

We also need to verify the stated conditions on the true world $x_0 = (x_{01}, \dots, x_{0n}) \in \mathbf{T}$ that make it possible for \mathcal{A} to have full knowledge acquisition about p . In view of Definition 2.3, we must verify that

$$(85) \quad \sup_{\boldsymbol{\lambda}} \mathbf{P}(B_{\epsilon}(x_0); \boldsymbol{\lambda}) = 1$$

for any ball $B_{\epsilon}(x_0)$ of radius $\epsilon > 0$ surrounding x_0 . Since each marginal density \mathbf{P}_i in (45) is monotone, it is clear that (85) holds only if for each $i = 1, \dots, n$, either $x_{0i} = 0$ or $x_{0i} = 1$, with the maximum in (85) being attained in the limit where $\lambda_i \rightarrow -\infty$ if $x_{0i} = 0$ and $\lambda_i \rightarrow \infty$ if $x_{0i} = 1$ respectively. \square

PROOF OF THEOREM 4.3. Assume without loss of generality that p is true (the proof is analogous when p is false) and that the supremum norm $d(x, y) = \max_{1 \leq i \leq n/2} |x_i - y_i|$ is used as a distance between the elements of \mathcal{X} . Since, by assumption, $x_0 = (x_{01}, \dots, x_{0, n/2})$ is an inner point of \mathbb{T} , we can choose $\varepsilon > 0$ so small that the ball of radius ε around x_0 is included in \mathbb{T} , i.e.

$$(86) \quad B(x_0, \varepsilon) = \times_{i=1}^{n/2} [x_{0i} - \varepsilon, x_{0i} + \varepsilon] \subset \mathbb{T}.$$

It follows from (44), (47) and (86) that

$$\mathbf{P}(\mathbb{T}) \geq \mathbf{P}(B(x_0, \varepsilon)) = \prod_{i=1}^{n/2} \frac{\int_{x_{0i}-\varepsilon}^{x_{0i}+\varepsilon} e^{\lambda_{2i-1}t + \lambda_{2i}t^2} dt}{\int_0^1 e^{\lambda_{2i-1}t + \lambda_{2i}t^2} dt} \rightarrow 1,$$

where the last limit holds if the components of λ are chosen pairwise, for each feature $i = 1, \dots, n/2$, so that

$$\begin{aligned} \lambda_{2i-i} &\rightarrow \infty, \\ \lambda_{2i} &\rightarrow -\infty, \\ \lambda_{2i-1} + 2x_{0i}\lambda_{2i} &= 0. \end{aligned}$$

The last displayed equation implies that agent \mathcal{A} 's posterior density $\mathbf{P}_i(x_i)$ for coordinate x_i is maximized at x_{0i} and converges weakly to a point mass at x_{0i} . Together with the coordinatewise independence (44), this implies that \mathbf{P} converges weakly to a point mass at x_0 . \square

PROOF OF THEOREM 4.4. To $\mathcal{X} = [0, 1]^r$ we assign a uniform prior $\mathbf{P}_0(x) \equiv 1$. The finite partition $\mathcal{P} = \{A_1, \dots, A_n\}$ of \mathcal{X} corresponds to n feature indicator functions (48), and the posterior distribution is given by (49), with

$$(87) \quad p_i = \frac{\mu_i}{|A_i|} = \frac{e^{\lambda_i}}{|A_1|e^{\lambda_1} + \dots + |A_n|e^{\lambda_n}} \propto e^{\lambda_i}.$$

Here $\mu_i = \mathbf{P}(A_i)$ is agent \mathcal{A} 's belief about the value of feature i , p_i is the value of $\mathbf{P}(x)$ on A_i , and $|A_i| = \nu(A_i)$ is the Lebesgue measure of A_i . Since the feature functions f_i are linearly dependent (51), without loss of generality we may choose λ so that the last proportionality of (87) is an equality.

In order to construct the posterior distribution from a recursively partitioned binary tree, the sets A_i must be r -dimensional rectangles with sides parallel to the r coordinate axes. In more detail, we make use of a binary tree

$$\mathcal{T} = \{t_1, \dots, t_{2n-1}\} = \mathcal{T}_1 \cup \mathcal{T}_2$$

with $2n - 1$ nodes, of which those in $\mathcal{T}_1 = \{t_1, \dots, t_n\}$ are leaves, those in $\mathcal{T}_2 = \{t_{n+1}, \dots, t_{2n-1}\}$ are inner nodes, and t_{2n-1} is the root of the tree. In particular, A_i and p_i are, respectively, a region and a probability weight associated with leaf node t_i , for $i = 1, \dots, n$. Each node $t \in \mathcal{T}$ is represented as a binary sequence

$$(88) \quad t = (m_{t1}, \dots, m_{th_t})$$

of length h_t , where h_t is the height of t , i.e. the number of edges of the path from the root t_{2n-1} to t . Edge number k of this path corresponds to a left turn (right turn) if $m_{tk} = 0$ ($m_{tk} = 1$). The height of the whole tree is the maximal height

$$h = \max(h_{t_1}, \dots, h_{t_n})$$

of all leaf nodes, and the tree is balanced if $h = h_{t_i}$ for all leaf nodes. For each $t \in \mathcal{T}$, we define the parental set

$$\text{pa}(t) = \begin{cases} \{(m_{t_1}, \dots, m_{t, h_t-1})\}, & t \neq t_{2n-1}, \\ \emptyset, & t = t_{2n-1}, \end{cases}$$

and the offspring set

$$\text{off}(t) = \begin{cases} \emptyset, & t \in \mathcal{T}_1, \\ \{\text{ch}_0(t), \text{ch}_1(t)\}, & t \in \mathcal{T}_2, \end{cases}$$

where the two children of an inner node are defined through $\text{ch}_l(t) = (m_{t_1}, \dots, m_{t_{h_t}}, l)$ for $l = 0, 1$. We also define $t(k) = (m_{t_1}, \dots, m_{t_k})$ as the $(h_t - k)$ -fold parent of t for $k = 0, \dots, h_t - 1$, with $t(0) = t_{2n-1}$ and $t(h_t - 1) = \text{pa}(t)$. The set A_i and the probability weight p_i are built recursively along the path that connects the root t_{2n-1} with $t_i \in \mathcal{T}_1$. In order to describe this construction in more detail, we associate with each inner node $t \in \mathcal{T}_2$ a splitting coordinate $j_t \in \{1, \dots, r\}$, a splitting point $a_t \in (0, 1)$ and a splitting probability $q_t \in (0, 1)$. When $t \in \mathcal{T}_2$ is branched to have two offspring $\text{ch}_0(t)$ and $\text{ch}_1(t)$, we let

$$B_t = \{x \in \mathcal{X}; x_{j_t} \geq a_t\}$$

be the splitting set associated with the right turn $\text{ch}_1(t)$, and its complement B_t^c the set that corresponds to the left turn $\text{ch}_0(t)$, where x_{j_t} is the j_t -th coordinate of $x \in \mathcal{X}$. Then, for each leaf node $t_i \in \mathcal{T}_1$, put

$$(89) \quad \mu_i = \prod_{k=1}^{h_{t_i}} \left[q_{t_i(k-1)}^{m_{t_i k}} (1 - q_{t_i(k-1)})^{1-m_{t_i k}} \right],$$

$$(90) \quad A_i = \bigcap_{k=1}^{h_{t_i}} \left[\mathbb{1} \{m_{t_i(k-1)} = 1\} B_{t_i(k-1)} + \mathbb{1} \{m_{t_i(k-1)} = 0\} B_{t_i(k-1)}^c \right],$$

and

$$(91) \quad |A_i| = \prod_{k=1}^{h_{t_i}} \left[(1 - a_{t_i(k-1)})^{m_{t_i k}} a_{t_i(k-1)}^{1-m_{t_i k}} \right].$$

From (87), (89) and (91), it follows that, without loss of generality, the parameters λ_i of the Gibbs distribution \mathbf{P} can be chosen as

$$(92) \quad \begin{aligned} \lambda_i &= \log p_i \\ &= \sum_{k=1}^{h_{t_i}} \left[m_{t_i k} \log q_{t_i(k-1)} + (1 - m_{t_i k}) \log (1 - q_{t_i(k-1)}) \right] \\ &\quad - \sum_{k=1}^{h_{t_i}} \left[m_{t_i k} \log (1 - a_{t_i(k-1)}) + (1 - m_{t_i k}) \log q_{t_i(k-1)} \right] \\ &= \sum_{k=1}^{h_{t_i}} \left[m_{t_i k} \log \frac{q_{t_i(k-1)}}{1 - a_{t_i(k-1)}} + (1 - m_{t_i k}) \log \frac{1 - q_{t_i(k-1)}}{a_{t_i(k-1)}} \right]. \end{aligned}$$

If the feature functions f_i are fixed (that is, if j_t and a_t are fixed for all $t \in \mathcal{T}_1$), then agent \mathcal{A} chooses splitting probabilities q_t for all $t \in \mathcal{T}_1$ in order to compute the feature coefficients (92) of his posterior.

Since \mathcal{P} is a partition of \mathcal{X} ,

$$\max_{1 \leq i \leq n} |A_i| \geq \frac{1}{n}.$$

Moreover, since each A_i is a rectangle, its diameter satisfies

$$\text{diam}(A_i) = \max\{d(x, y); x, y \in A_i\} \geq |A_i|^{1/r},$$

where $d(x, y) = \max_{1 \leq j \leq r} \|x_j - y_j\|$ is the supremum norm in $[0, 1]^r$. From the last two displayed equations, we find that

$$(93) \quad 2\varepsilon = \max_{1 \leq i \leq n} \text{diam}(A_i) \geq \frac{1}{n^{1/d}} \geq \frac{1}{2^{h/d}},$$

where the last inequality follows from $n \leq 2^h$, with equality for balanced trees. Since all $A_i \in \mathcal{P}$ are rectangles, and the posterior (50) is constant on each A_i , we deduce from (93) that $x_0 \in \mathcal{X}$ can be chosen so that

$$(94) \quad \mathbf{P}(B(x_0, \varepsilon)) < 1.$$

We see from (94) that $n \rightarrow \infty$ is a necessary condition in order to guarantee asymptotic full knowledge of x_0 , i.e., $\mathbf{P}(B(x_0, \varepsilon)) \rightarrow 1$ as $n \rightarrow \infty$ for each $\varepsilon > 0$. \square

PROOF OF PROPOSITION 4.2. Observe that \mathcal{G}_A in (55) is the collection of sets B such that either B or $[0, 1] \setminus B$ is a subset of A . The difference from Billingsley's example is that the set A is now fixed, not an arbitrary countable subset of $[0, 1]$. Since \mathcal{G}_A is generated by a countable collection (8) of sets, we apply (10) to conclude that the probability measure of agent \mathcal{A} must satisfy

$$(95) \quad \mathbf{P}(x) = p_0 + \sum_{i=1}^{\infty} p_i \delta_{x_i}(x)$$

for some non-negative numbers p_i satisfying $\sum_{i=0}^{\infty} p_i = 1$. That is, the belief of \mathcal{A} about x_0 is a mixture of ignorance (a uniform density with weight p_0) and a belief that is supported on A . This is to say that data D supply \mathcal{A} with information that x_0 either belongs to the set A or it can be any other element of $[0, 1]$. Consider, without loss of generality, the proposition

$$p : x_0 \text{ belongs to the set } [0.5, 1].$$

It follows from (2) that $f_p(x) = \mathbb{1}_T(x)$, with $T = [0.5, 1]$. Although $T \notin \mathcal{G}_A$ and f_p is not measurable with respect to \mathcal{G}_A , if p is true and $A \cap T \neq \emptyset$ it is still possible for \mathcal{A} to fully learn p (when $p_0 = 0$ and $p_i = 0$ for all $x_i \notin T$ in (95)) and additionally acquire full knowledge about p (if also $x_0 = x_i \in A \cap T$ and $p_i = 1$). Analogously, if p is false and $A \cap T^c \neq \emptyset$, it is possible for \mathcal{A} to learn p fully and additionally acquire full knowledge about p , if also $x_0 \in A \cap T^c$. However, since \mathbf{P} is constructed as an infinite sum, it is not possible to express (95) in terms of a Gibbs distribution. This proves the first part of the proposition.

To prove the second part, consider the smaller σ -field (56) constructed from the finite set $A_n = \{x_1, \dots, x_n\}$. It follows from (10) that the posterior belief of \mathcal{A} must satisfy

$$(96) \quad \mathbf{P}(x) = p_0 + \sum_{i=1}^n p_i \delta_{x_i}(x),$$

for some non-negative numbers p_i such that $\sum_{i=0}^n p_i = 1$. The distribution in (96) can be approximated by a Gibbs distribution (18) with n features, as follows: Assume $0 < x_i < 1$ for $i = 1, \dots, n$ and choose $\delta > 0$ so small that all $A_i = [x_i - \delta/2, x_i + \delta/2]$ are disjoint. Then introduce the spiky feature functions

$$(97) \quad f_i(x) = f_i(x; \delta) = \mathbb{1}_{A_i}(x) \log \delta^{-1}$$

for $i = 1, \dots, n$. Let also $C = [0, 1] \setminus \cup_{i=1}^n A_i$. It follows from (18) that the Gibbs distribution based on features (97) is given by

$$\begin{aligned}
 \mathbf{P}(x) &= Z_{\lambda}^{-1} \left[\mathbb{1}_C(x) + \delta^{-1} \sum_{i=1}^n \mathbb{1}_{A_i}(x) e^{\lambda_i} \right] \\
 (98) \quad &= p_0(\delta) \mathbb{1}_C(x) + \delta^{-1} \sum_{i=1}^n p_i(\delta) \mathbb{1}_{A_i}(x) \\
 &\xrightarrow{\mathcal{L}} p_0 + \sum_{i=1}^n p_i \delta_{x_i}(x),
 \end{aligned}$$

where $p_0(\delta) = 1/Z_{\lambda}$, $p_i(\delta) = e^{\lambda_i}/Z_{\lambda}$ for $i = 1, \dots, n$, and $Z_{\lambda} = 1 - n\delta + \sum_{i=1}^n e^{\lambda_i}$. The last step of (98) refers to weak convergence as $\delta \rightarrow 0$, with

$$\begin{aligned}
 (99) \quad p_0 &= \lim_{\delta \rightarrow 0} p_0(\delta) = \frac{1}{1 + \sum_{j=1}^n e^{\lambda_j}}, \\
 p_i &= \lim_{\delta \rightarrow 0} p_i(\delta) = \frac{e^{\lambda_i}}{1 + \sum_{j=1}^n e^{\lambda_j}}, \quad i = 1, \dots, n.
 \end{aligned}$$

□

PROOF OF PROPOSITION 5.1. From the asymptotic theory of maximum likelihood estimates, we find that the estimate $\hat{\lambda}$ of λ in (59) is asymptotically normally distributed

$$(100) \quad \sqrt{m} (\hat{\lambda} - \lambda) \xrightarrow{\mathcal{L}} N(0, \mathbf{J}^{-1})$$

as $m \rightarrow \infty$. Next we insert (59) into (62) and perform a second order Taylor expansion of $\text{Bias}(\mathbf{T}; \lambda, \hat{\lambda})$ with respect to $\hat{\lambda}$ around λ . Taking the expectation of this Taylor expansion, with respect to random variations in $\tilde{\mathbf{D}}$, and making use of (100), we finally obtain (63). □

REFERENCES

- [1] BARBIER, J. (2020). High-dimensional inference: a statistical mechanics perspective. *Ithaca: Viaggio nella Scienza* **XVI**.
- [2] BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd. ed. Wiley.
- [3] BOYD, S. and VANDENBERGUE, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.
- [4] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees. The Wadsworth statistics/probability series*. Chapman and Hall/CRC, Boca Raton.
- [5] CELIS, L. E., KESWANI, V. and VISHNOI, N. K. (2020). Data preprocessing to mitigate bias: A maximum entropy based approach. *International Conference on Machine Learning* 1349-1359.
- [6] CHERIAN, J. J., GIBBS, I. and CANDES, E. J. (2024). Large language model validity via enhanced conformal prediction methods. *arXiv:2406.09714v2*.
- [7] COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, Second ed. Wiley.
- [8] DAVIES, P. (2019). *The Demon in the Machine*. Allen Lane, Great Britain.
- [9] DEMBSKI, W. A. and MARKS II, R. J. (2009). Conservation of Information in Search: Measuring the Cost of Success. *IEEE Transactions Systems, Man, and Cybernetics - Part A: Systems and Humans* **5** 1051-1061. <https://doi.org/10.1109/TSMCA.2009.2025027>
- [10] DEMBSKI, W. A. and MARKS II, R. J. (2010). The Search for a Search: Measuring the Information Cost of Higher Level Search. *Journal of Advanced Computational Intelligence and Intelligent Informatics* **14** 475-486. <https://doi.org/10.20965/jaciii.2010.p0475>
- [11] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, Cham.

- [12] DÍAZ-PACHÓN, D. A. and HÖSSJER, O. (2022). Assessing, testing and estimating the amount of fine-tuning by means of active information. *Entropy* **24** 1323. <https://doi.org/10.3390/e24101323>
- [13] DÍAZ-PACHÓN, D. A., HÖSSJER, O. and MARKS II, R. J. (2021). Is Cosmological Tuning Fine or Coarse? *Journal of Cosmology and Astroparticle Physics* **2021** 020. <https://doi.org/10.1088/1475-7516/2021/07/020>
- [14] DÍAZ-PACHÓN, D. A., HÖSSJER, O. and MARKS II, R. J. (2023). Sometimes size does not matter. *Foundations of Physics* **53** 1. <https://doi.org/10.1007/s10701-022-00650-1>
- [15] DÍAZ-PACHÓN, D. A., HÖSSJER, O. and MATTHEW, C. (2024). Is It Possible to Know Cosmological Fine-tuning? *The Astrophysical Journal Supplement Series* **271** 56. <https://doi.org/10.3847/1538-4365/ad2c88>
- [16] DÍAZ-PACHÓN, D. A. and MARKS II, R. J. (2020). Active Information Requirements for Fixation on the Wright-Fisher Model of Population Genetics. *BIO-Complexity* **2020** 1-6. <https://doi.org/10.5048/BIO-C.2020.4>
- [17] DÍAZ-PACHÓN, D. A. and RAO, J. S. (2021). A simple correction for COVID-19 sampling bias. *Journal of Theoretical Biology* **512** 110556. <https://doi.org/10.1016/j.jtbi.2020.110556>
- [18] DÍAZ-PACHÓN, D. A., SÁENZ, J. P. and RAO, J. S. (2020). Hypothesis testing with active information. *Statistics & Probability Letters* **161** 108742. <https://doi.org/10.1016/j.spl.2020.108742>
- [19] DÍAZ-PACHÓN, D. A., SÁENZ, J. P., RAO, J. S. and DAZARD, J.-E. (2019). Mode hunting through active information. *Applied Stochastic Models in Business and Industry* **35** 376-393. <https://doi.org/10.1002/asmb.2430>
- [20] DUDÍK, M. (2007). Maximum Entropy Density Estimation with Generalized Regularization and an Application to Species Distribution Modeling. *Journal of Machine Learning Research* **8** 1217-1260.
- [21] GETTIER, E. L. (1963). Is Justified True Belief Knowledge? *Analysis* **23** 121-123. <https://doi.org/10.2307/3326922>
- [22] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second ed. Springer Science, New York.
- [23] HÖSSJER, O., DÍAZ-PACHÓN, D. A., CHEN, Z. and RAO, J. S. (2024). An Information Theoretic Approach to Prevalence Estimation and Missing Data. *IEEE Transactions on Information Theory* **70** 3567-3582. <https://doi.org/10.1109/TIT.2023.3327399>
- [24] HÖSSJER, O., DÍAZ-PACHÓN, D. A. and RAO, J. S. (2022). A Formal Framework for Knowledge Acquisition: Going beyond Machine Learning. *Entropy* **24** 1469. <https://doi.org/10.3390/e24101469>
- [25] ICHIKAWA, J. J. and STEUP, M. (2018). The Analysis of Knowledge. In *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.) Metaphysics Research Lab, Stanford University, Stanford.
- [26] LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, Second ed. Springer.
- [27] LEVIN, D. A. and PERES, Y. (2017). *Markov Chains and Mixing Times*. American Mathematical Society, Providence.
- [28] LI, Z., ZHU, H., LU, Z. and MING, Y. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *Proceedings of the 2023 Conference on Natural Language Processing* 10443-10461.
- [29] LIU, T., DÍAZ-PACHÓN, D. A., RAO, J. S. and DAZARD, J.-E. (2023). High Dimensional Mode Hunting Using Pettiest Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45** 4637-4649. <https://doi.org/10.1109/TPAMI.2022.3195462>
- [30] MCMILLEN, P., WALKER, S. I. and LEVIN, M. (2022). Information Theory as an Experimental Tool for Integrating Disparate Biophysical Signaling Modules. *International Journal of Molecular Sciences* **23** 9580. <https://doi.org/10.3390/ijms23179580>
- [31] MOORE, D. G., WALKER, S. I. and LEVIN, M. (2017). Cancer as a disorder of patterning information: Computational and biophysical perspectives on the cancer problem. *Convergent Science Physical Oncology* **3** 043001. <https://doi.org/10.1088/2057-1739/aa8548>
- [32] MONTAÑEZ, G. D. (2017). The famine of forte: Few search problems greatly favor your algorithm. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 477-482. <https://doi.org/10.1109/SMC.2017.8122651>
- [33] MONTAÑEZ, G. D., BASHIR, D. and LAUW, J. (2021). Trading Bias for Expressivity in Artificial Learning. In *Agents and Artificial Intelligence* (A. P. ROCHA, L. STEELS and J. VEN DEN HERIK, eds.) 332-353. Springer, Cham. https://doi.org/10.1007/978-3-030-71158-0_16
- [34] MONTAÑEZ, G. D., HAYASE, J., LAUW, J., MACIAS, D., TRIKHA, A. and VENDEMIATTI, J. (2019). The Futility of Bias-Free Learning and Search. In *2nd Australasian Joint Conference on Artificial Intelligence (AI 2019)* (J. Liu and J. Bailey, eds.) 277-288. Springer, Cham. https://doi.org/10.1007/978-3-030-35288-2_23
- [35] PIETRA, S. D., PIETRA, V. D. and LAFFERTY, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** 380-393. <https://doi.org/10.1109/34.588021>

- [36] RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
- [37] SCHWITZGEBEL, E. (2021). Belief. In *The Stanford Encyclopedia of Philosophy* winter 2021 ed. (E. N. Zalta, ed.) Metaphysics Research Lab, Stanford University, Stanford.
- [38] THORVALDSEN, S. and HÖSSJER, O. (2023). Estimating the Information Content of Genetic Sequence Data. *Journal of the Royal Statistical Society Series C: Applied Statistics* **72** 1310-1338. <https://doi.org/10.1093/jrsssc/qlad062>
- [39] THORVALDSEN, S. and HÖSSJER, O. (2024). Use of directed quasi-metric distances for quantifying the information of gene families. *BioSystems* **243** 105256. <https://doi.org/10.1016/j.biosystems.2024.105256>
- [40] WALKER, S. I. and DAVIES, P. (2013). The algorithmic origins of life. *J. R. Society Interface* **10** 20120869. <https://doi.org/10.1098/rsif.2012.0869>
- [41] WIBRAL, M., LIZIER, J. T. and PRIESEMANN, V. (2014). How to measure local active information storage in neural systems In *8th Conf. of the European Study Group on Cardiovascular Oscillations* 131-132. <https://doi.org/10.1109/ESGCO.2014.6847554>
- [42] WIBRAL, M., LIZIER, J. T. and PRIESEMANN, V. (2015). Bits from Brains for Biologically Inspired Computing. *Frontiers in Robotics and AI* **2**. <https://doi.org/10.3389/frobt.2015.00005>
- [43] WIBRAL, M., LIZIER, J. T., VÖGLER, S., PRIESEMANN, V. and GALUSKE, R. (2014). Local active information storage as a tool to understand distributed neural information processing. *Frontiers in Neuroinformatics* **8**. <https://doi.org/10.3389/fninf.2014.00001>
- [44] ZDEBOROVÁ, L. and KRZAKALA, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Advances in Physics* **65** 453-552. <https://doi.org/10.1080/00018732.2016.1211393>
- [45] ZHOU, L., DÍAZ-PACHÓN, D. A., ZHAO, C., RAO, J. S. and HÖSSJER, O. (2023). Correcting prevalence estimation for biased sampling with testing errors. *Statistics in Medicine* **42** 4713-4737. <https://doi.org/10.1002/sim.9885>