Learning as Hypothesis Testing: Learning Conditional and Probabilistic Information

Jonathan Vandenburgh[*]

February 25, 2021

Abstract

Complex constraints like conditionals ('If A, then B') and probabilistic constraints ('The probability that A is p') pose problems for Bayesian theories of learning. Since these propositions do not express constraints on outcomes, agents cannot simply conditionalize on the new information. Furthermore, a natural extension of conditionalization, relative information minimization, leads to many counterintuitive predictions, evidenced by the sundowners problem and the Judy Benjamin problem. Building on the notion of a 'paradigm shift' and empirical research in psychology and economics, I argue that the model of hypothesis testing can explain how people learn complex, theory-laden propositions like conditionals and probability constraints. Theories are formalized as probability distributions over a set of possible outcomes and theory change is triggered by a constraint which is incompatible with the initial theory. This leads agents to consult a higher order probability function, or a 'prior over priors,' to choose the most likely alternative theory which satisfies the constraint. The hypothesis testing model is applied to three examples: learning a simple probabilistic constraint involving coin bias, the sundowners problem for conditional learning, and the Judy Benjamin problem for learning conditional probability constraints.

Suppose you are rolling what you believe to be a fair die, but then come to learn the conditional sentence 'If you roll an odd number, the outcome will be three.' Given this new information, how should you change your credences in each of the six possible outcomes? It seems, at minimum, that you must rule out 1 and 5 as possible outcomes, but the new constraint is consistent with many hypotheses about the die. Perhaps all faces of the die are labeled 3, or perhaps the die has only even numbers on it.

The dominant approach to this kind of learning problem assumes that agents incorporate the new constraint in a way which minimizes the relative difference in information between the posterior and the prior credences (Williams, 1980; Diaconis and Zabell, 1982). However, this approach often yields counterintuitive results. In the above case, information minimization requires that one eliminate the outcomes of 1 and 5 and redistribute the likelihoods of the other outcomes in a way which is proportional to one's initial beliefs, so that the chances of rolling each number from $\{2, 3, 4, 6\}$ become $\frac{1}{4}$. This, however, seems irrational: if the die is balanced and six-sided with one number per side, there is no way to realize a one in four chance of landing on a particular number.

[*] Comments are welcome at jonathanvandenburgh2021@u.northwestern.edu.

The motivation for information minimization, inspired by Bayesian learning, is that agents learn things about the world in a way which is gradual and cumulative, building on prior beliefs. This conservative approach is also at the heart of other, non-probabilistic approaches to learning, such as the AGM model (Alchourrón et al., 1985; Gärdenfors, 1988).[1] However, some philosophers and social scientists have argued that learning can be disruptive, occasionally requiring radical changes to prior beliefs. A particularly strong version of this idea appears in Kuhn's (2012) theory of paradigm shifts, where the foundational beliefs of an epistemic community undergo radical revision, as when a generation of scientists had to adapt to the evidence supporting the heliocentric model. Versions of this disruptive approach to learning have also appeared in empirical work in psychology and economics. One approach to the psychology of concepts suggests that children learn concepts by proposing and testing theories of concept membership (Carey and Spelke, 1994; Gopnik, 1996). For example, learning that dolphins are mammals rather than fish does not proceed simply by adding dolphins to the list of things that are mammals, but requires more substantial changes to the theory of what makes something a mammal versus a fish. Furthermore, experiments on strategic behavior in economics suggest that people develop theories about how their opponents are strategizing and, when the opponent's behavior differs significantly from the predictions of the theory, completely abandon the hypothesis in favor of a new one (Salmon, 2004; Young, 2004).

One way of modeling disruptive learning is through the framework of hypothesis testing, where agents replace one hypothesis or theory with an alternative selected from a hypothesis space according to a higher-order probability function, or a 'prior over priors' (Young, 2004; Griffiths and Tenenbaum, 2009; Ortoleva, 2012). In this paper, I argue that the model of hypothesis testing can explain cases of learning complex information. The main cases studied are cases where people learn conditional information ('If A, then B') and probabilistic information ('The probability that A is p'), including conditional probability constraints ('The probability of B given A is p'), though the model is applicable beyond these cases. Throughout, I argue that the hypothesis testing model offers better predictions than the information minimization model, addressing challenging cases of learning such as the sundowners problem (Douven, 2012) and the Judy Benjamin problem (Van Fraassen, 1981). This theory is related to some recent work on higher-order probability spaces for the semantics of complex propositions, particularly probabilistic modals (Yalcin, 2012; Goldstein, 2020), but differs by addressing problems in the literature on Bayesian learning and by offering specific strategies for producing higher-order probability spaces.

The paper is organized as follows. In §1, I introduce Bayesian learning for simple constraints which restrict the space of possible outcomes. In §2, I characterize a set of theory-level constraints which cannot be incorporated in the framework of Bayesian learning. In §3, I introduce the hypothesis testing model for learning theory-level constraints incompatible with an initial theory. I argue that this theory involves two steps: (1) identifying a hypothesis space of alternative theories satisfying the new constraint and (2) choosing a new theory according to a higher-order probability function. I introduce some tools to simplify these steps and discuss the problem of underdetermination, where the posterior credences may be underdetermined by the prior credences and the constraint. The remainder of the paper applies the theory to three examples: a case of probabilistic information concerning coin bias (§4), the sundowners problem for learning conditional information (§5), and the Judy Benjamin

[1] For a criticism of the conservatism of the AGM model specifically, see Gillies (2006).

problem for learning conditional probability constraints (§6). In each case, the predictions are contrasted with those of the information minimization model.

1 Bayesian Learning

Suppose someone is rolling a die and expects the die to be fair, so that the possible outcomes are the numbers $\{1, ..., 6\}$, each with probability $\frac{1}{6}$ of occurring. Upon learning that the die landed on an odd number, the natural way to update beliefs is to restrict the set of possible outcomes to $\{1, 3, 5\}$, each with probability $\frac{1}{3}$. The new beliefs are formed by conditionalizing on the proposition that the die is odd: the space of possibilities is restricted to the odd numbers and the new probability distribution is given by $\Pr(\{n\}|\text{odd}) = \frac{\Pr(\{n\}\wedge\text{odd})}{\Pr(\text{odd})}$. This procedure of conditionalization defines Bayesian learning.

To discuss Bayesian learning more formally, we start by representing the agent's belief space as a set of possible outcomes Ω with a probability distribution π over Ω. When Ω is finite, the probability distribution π is simply an assignment $\pi(\omega) \in [0, 1]$ to each element ω of Ω such that $\sum_{\omega\in\Omega} \pi(\omega) = 1$.[2] The target propositions for learning are subsets of Ω, $A \subseteq \Omega$, so the set of propositions of interest is $\mathcal{A} = \mathcal{P}(\Omega)$, the set of subsets of Ω. Note that this resembles the set-up for possible world semantics, where Ω plays the role of the set of all possible worlds and propositions correspond to sets of possible worlds.

Since π is a probability distribution on Ω and Ω is finite, π gives a probability assignment to every subset of Ω, or to every proposition in $\mathcal{A}$.[3] This is just given by $\pi(A) = \sum_{\omega\in A} \pi(\omega)$. For any two subsets $A, B \in \mathcal{A}$, we can define the conditional probability $\pi(B|A) = \frac{\pi(A\wedge B)}{\pi(A)}$. Here, the logical operation $\wedge$ is determined by set-theoretic intersection $\cap$. This gives the probability that B is true conditional on A.

This allows us to define a procedure for Bayesian updating. When someone with belief space (Ω, π) learns a proposition $A \in \mathcal{A}$, they form a new probability distribution π_A over Ω by conditionalizing on A: $\pi_A(\omega) = \frac{\pi(\omega\wedge A)}{\pi(A)}$. Note that $\pi(\omega \wedge A)$ is 0 when $\omega \notin A$ and $\pi(\omega)$ when $\omega \in A$. This means that learning A restricts to just those outcomes where A is true, and, for outcomes where A is true, $\pi_A(\omega)$ stays proportional to the prior probability $\pi(\omega)$. This makes sense because we have only learned information about whether the outcome is in A, not about which outcomes in A are more or less likely. For any proposition B, the probability of B in the updated distribution is

$$\sum_{\omega\in B} \pi_A(\omega) = \frac{1}{\pi(A)} \sum_{\omega\in B} \pi(\omega \wedge A) = \frac{\pi(A \wedge B)}{\pi(A)},$$

which is just the conditional probability, so $\pi_A(B) = \pi(B|A)$.

To return to the example of rolling a die, assume that $\Omega = \{1, ..., 6\}$ and $\pi(\omega) = \frac{1}{6}$ for each ω. The proposition 'The die landed on an odd number' can be represented as the set $A = \{1, 3, 5\}$, where $\pi(A) = \sum_{\omega\in A} \pi(\omega) = \frac{1}{2}$. Bayesian updating on A leads to the posterior distribution where the new set of possible worlds is A and for each $\omega \in A$, $\pi_A(\omega) = \frac{\pi(\omega)}{\pi(A)} = \frac{1/6}{1/2} = \frac{1}{3}$. This yields the expected updated distribution, where the set of possibilities is restricted to the three odd numbers, each with probability $\frac{1}{3}$.

[2] In the infinite case, we require that there is a sigma algebra Σ representing the set of measurable subsets of Ω and that $\pi : \Sigma \to [0, 1]$ satisfies the following: (i) $\pi(\Omega) = 1$ and (ii) $\forall A_i \in \Sigma, \pi(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \pi(A_i)$.

[3] In the infinite case, we must restrict to $\mathcal{A} \subseteq \Sigma$.

2 Theories and Incompatible Constraints

In discussing Bayesian learning, we have modeled an agent's belief space as a probability distribution over a set of possibilities, (Ω, π). We can think of this probability distribution as the agent's *hypothesis* or *theory* about the world, representing which outcomes are possible and what their likelihoods are. For example, the agent's belief space from the previous section encodes the theory that a given die is fair, with the die having an equal probability of landing on each of the numbers 1 through 6. Similarly, we can think of other theories in terms of a probability distribution: a theory of human behavior in a situation, for example, would specify which actions are possible and how likely each action is.[4] Theories can also include other information, like causal information (that rain can cause changes in event plans, but not the other way around) or the definitions or structures of basic terms (like that fish are water-dwellers or that mammals have mammary glands).[5] While representing theories as probability distributions over a set of possibilities does not explicitly include this additional information, we will see how it can be incorporated in the next section and the example in §5.

Bayesian learning focuses on a specific kind of constraint an agent learns: propositions expressed as a subset of the sample space Ω. These constraints provide information about the outcome itself, such as that a die landed on an odd number or that certain actions are ruled out. However, not all constraints correspond to a subset of the sample space. Two examples of such constraints are conditional propositions like 'If the die lands on an odd number, then the die lands on 3' and propositions which express probabilistic constraints like 'The probability that the die lands on 3 is $\frac{1}{2}$.' Neither of these propositions introduce constraints on the outcome itself: the die landing 4, for example, does not settle whether either constraint is satisfied.

We can think of these propositions instead as providing constraints on the entire belief space (Ω, π). Typically, the truth values of conditionals (Kratzer, 1986, 2012) and probability operators (Yalcin, 2010) depend on the set of relevant possible alternatives to ω, which we can think of as the set Ω. For example, the conditional 'If the die lands on an odd number, then the die lands on 3' does not become true based on the number rolled; instead, it is true when the set of possibilities Ω satisfies the constraint that 3 is the only odd number one can roll. Similarly, the probabilistic constraint 'The probability that the die lands on 3 is $\frac{1}{2}$' is satisfied by any (Ω, π) where $\pi(\{3\}) = \frac{1}{2}$. For example, if a six-faced fair die has three 2s and three 3s on it, so $\Omega = \{2, 3\}$ and for each $\omega, \pi(\omega) = \frac{1}{2}$, then both the conditional and the probabilistic constraint are satisfied.

Conditional and probabilistic constraints are examples of *theory-level constraints* which restrict the global theory (Ω, π) rather than the outcome ω. Theory-level constraints include both constraints on the global set of possibilities Ω and constraints on the probability distribution π. The former case includes conditional constraints, as well as modal constraints about what is possible and necessary. While I do not discuss modal constraints in detail in this

[4] This is how hypotheses about opponent behavior are formalized in strategic games; see Foster and Young (2003); Young (2004).

[5] For some authors, the basic terms (ontology) and the kinds of relationships which are possible in the first place are encoded in more basic 'foundation or framework theories' rather than 'specific theories.' In this paper, I focus on specific theories rather than foundational theories. For more on this distinction, see Griffiths and Tenenbaum (2009) and Wellman and Gelman (1992).

paper, this category includes cases where people learn about new and unforeseen possibilities as well as new laws which hold of necessity in all possible worlds. Probabilistic constraints include the constraint discussed above, that $\pi(A) = p$ for some $A \subseteq \Omega$ and $p \in [0,1]$, as well as more complex constraints like the conditional probability constraint that $\pi(B|A) = q$ for $A, B \subseteq \Omega$ and $q \in [0,1]$.

One characteristic of theory-level constraints is that constraints which are not part of the initial theory are impossible to satisfy within the theory. This contrasts with constraints on outcomes, which, even when left open by the initial theory, are still possible to satisfy within the theory. Consider the case where one believes that a die is fair and learns the outcome-level constraint that the die lands on an odd number. This constraint is not part of the initial theory, but it is possible to constrain one's beliefs to satisfy the condition without revising the theory that the die is fair. On the other hand, if one learns the theory-level conditional constraint 'If the die lands odd, it lands 3,' it is impossible to satisfy this condition without abandoning the theory that the die is fair. Thus, theory-level constraints which are not satisfied on the initial theory appear similar to the kinds of constraints Gopnik (1996, p. 496) describes as requiring radical revision to an agent's theory: 'If we are committed to the theory such violations should strike us, not only as surprising, but as being impossible and unbelievable in an important and strong way.' In dynamic semantics, these constraints are considered tests of information states rather than propositional updates: they can provide a test for whether an epistemic state is true or false, but they do not offer a direct way of updating an information state as other propositions do (Veltman, 1996).

So far, I have argued that we can think of probability distributions over a set of outcomes as theories and constraints with likelihood zero on an initial theory as incompatible with the theory. Incompatible constraints cannot be learned through Bayesian updating and the incompatibility with an initial theory is what will motivate the need for theory revision discussed in the next section. It is important to note that psychologists and philosophers of science are often interested in a more robust notion of theory incompatibility than used here. Some theories, called framework or foundational theories (Wellman and Gelman, 1992; Griffiths and Tenenbaum, 2009), encode a basic ontology of concepts and the relations which are possible between these concepts. The kind of incompatibility relevant for these theories is *incommensurability*, which occurs when a constraint is not only impossible to realize in an initial theory, but is impossible to adequately state in the theory. For example, a child learning 'Clocks are not alive' may not even be able to express the new proposition in their prior vocabulary, which associates life with movement and includes things like clocks and the sun as alive (Carey, 2009). While learning incommensurable constraints is an important, though controversial, case of theory revision, this paper focuses on the broader class of incompatible constraints, which includes the conditional and probabilistic constraints of interest.

3 The Hypothesis Testing Model

We are interested in how agents learn from constraints which are incompatible with their initial theory or hypothesis about the world. Recall that Bayesian learning provides a procedure where an agent with beliefs (Ω, π) updates their beliefs upon learning a proposition A to (A, π_A). We would similarly like a procedure which takes as input a belief state (Ω, π) and a constraint C incompatible with this belief state and outputs a new belief state (Ω_C, π_C) which satisfies the constraint C and represents a reasonable deviation from the original beliefs.

Since incompatible constraints are not expressible as a subset of the original sample space Ω and have zero likelihood on the initial theory, agents cannot use Bayesian updating to arrive at a new belief space. In philosophy, the most popular approach to learning incompatible constraints involves finding a new probability distribution which satisfies the constraint C while deviating as little as possible from the prior beliefs. A common measure of how much the new beliefs differ from the old beliefs is the Kullback-Leibler divergence, which is designed to measure the amount of information learned in going from the prior to the posterior (Williams, 1980). Since Bayesian updating minimizes the Kullback-Leibler divergence, this seemingly represents a natural generalization of Bayesian learning. This procedure has been applied to both conditional and probabilistic constraints: for conditional constraints, this amounts to learning the material conditional (Douven and Romeijn, 2011; Eva et al., 2020), and for probabilistic constraints, this amounts to using Jeffrey's (1990) probability kinematics.

However, there is good reason to suspect that constraints which are incompatible with a theory should not always lead to minimal changes. Unlike the constraints from Bayesian learning, incompatible constraints require abandoning the initial theory of the world and settling on a new one. Instead of trying to incorporate these constraints smoothly, it makes sense to propose that this kind of learning instead involves a 'paradigm shift,' where the agent must go back to the drawing board and find a new theory which can explain the evidence. This motivates a hypothesis testing model, where an agent starts with an initial theory $h = (\Omega, \pi)$ and, upon learning an incompatible constraint C, chooses the most appealing alternative theory from a suitable *hypothesis space* $\mathcal{H}$ of possible alternative hypotheses. This model for learning incompatible constraints receives empirical support in psychology and economics. In psychology, theory revision is proposed as a theory of concepts, where agents categorize objects into groups by developing and revising theories about what it means to be a member of the kind (Gopnik, 1996; Carey and Spelke, 1994). In economics, hypothesis testing is used to explain how people update their theories of opponent behavior in strategic interaction, predicting that agents radically revise their theories when confronted with sufficient counterevidence (Salmon, 2004; Foster and Young, 2003; Young, 2004).

We can break the hypothesis testing model down into two steps. First, the agent finds an appropriate hypothesis space $\mathcal{H}$ consisting of possible theories, $\mathcal{H} = \{(\Omega_i, \pi_i)\}$, and restricts to the hypotheses satisfying the constraint C, $\mathcal{H}_C$. Second, the agent chooses a new hypothesis $h' = (\Omega_C, \pi_C)$ from $\mathcal{H}_C$. I discuss these two steps in greater detail below.

The difficulty of finding a suitable, manageably sized hypothesis space is likely the reason why the hypothesis testing model is given little attention, when mentioned at all, as an alternative to relative information minimization (Diaconis and Zabell, 1982; Halpern, 2017). Consider again the case of a six-sided die. If possible theories about the bias of the die include all possible probability distributions over the set of outcomes $\{1, ..., 6\}$, the space of possible theories is not only infinite, but is a five dimensional subset of $[0, 1]^6 \subset \mathbb{R}^6$. If learning a new constraint like 'If the die lands odd, it lands on 3' requires that the agent choose a new hypothesis from the set of possible theories satisfying the constraint with no conditions on distance from the prior, the set of hypotheses satisfying the constraint is infinitely large (in this case, a three dimensional subset of $[0, 1]^6$) and there appears to be no prediction for the agent's posterior credences.

However, there are many ways to simplify the hypothesis space for a learning problem such that the hypothesis testing model can offer predictions. One common method is the use of parametric models (Casella and Berger, 2002), where the hypothesis space consists of a set of parameter values which determine the full probability distribution. Consider an example: suppose an agent is trying to predict the height distribution of a population based on the height measurements from a small sample set. This agent is unlikely to consider all possible probability distributions consistent with their sample. Instead, the agent is likely to assume that the distribution has a certain form, such as the form of a normal distribution. Thus, the only hypotheses the agent considers are those built from normal distributions, so each theory in the hypothesis space is completely determined by two parameters, the estimated population mean μ and the estimated variation in heights σ. Using a parametric model in this way, the unmanageable hypothesis space of all possible height distributions is reduced to a smaller set of theories about the height distribution governed by two parameters. This smaller hypothesis space can offer predictions for learning: if we know the sample measurements the agent had access to, we can get a unique prediction for the agent's beliefs about the full height distribution. §4 uses a parametric model of the hypothesis space to offer predictions for learning probabilistic information in a simple coin bias case.

Another common approach to simplifying the hypothesis space involves using causal models (Rehder, 2003; Griffiths and Tenenbaum, 2009). Causal models impose constraints on how variables are related to each other, reducing the number of possible hypotheses. Causal models can simplify the hypothesis space by ruling out theories with inappropriate causal relations, like those where the outcomes of games of chance are causally linked with the weather. Additionally, when learning a new constraint, causal models can be used to delineate different hypotheses about the explanatory or causal status of the new constraint relative to one's prior beliefs. As we will see for conditional learning in §5, learning a conditional 'If A, then B' can lead to different causal hypotheses about the relationship between A and B, such as the hypothesis that A causes B or the hypothesis that B causes A, which adds enough structure to the learning problem to offer predictions.

Using the appropriate tools to build the hypothesis space $\mathcal{H}$ is what allows the hypothesis testing model to offer predictions. This requires figuring out how agents approach a specific learning problem and how this can be represented mathematically in a hypothesis space. Unlike the information minimization approach, this procedure depends on contextual features of the learning problem: while this requires more work for successful modeling, the goal of this paper is to argue that we can successfully find appropriate hypothesis spaces for a variety of learning problems and that doing so offers more compelling predictions than the information minimization model does.

Once we have the space $\mathcal{H}$ and have restricted to hypotheses which satisfy the constraint, $\mathcal{H}_C$, we can focus on the problem of choosing a new hypothesis from $\mathcal{H}_C$. In some cases, especially with parametric models as in §4, $\mathcal{H}_C$ will only have one element: in this case, the new theory h' will be the unique element from $\mathcal{H}_C$. In other cases, however, $\mathcal{H}_C$ will have multiple options for new beliefs consistent with the constraint, even with a simpler hypothesis space. In this case, we assume that the agent chooses the most likely hypothesis from $\mathcal{H}_C$. We can quantify likelihood with a probability distribution: we assume that the agent has a probability distribution Pr over $\mathcal{H}_C$ and chooses the new theory which is most likely according to Pr.[6] In many cases, Pr will offer a clear prediction for the new theory: in the example in

[6] We could assume that Pr is induced on $\mathcal{H}_C$ through Bayesian updating on a prior distribution over the entire hypothesis space $\mathcal{H}$, which is reasonable in many examples. However, in many cases (as in §4), $\mathcal{H}_C$ has zero probability as a subset of $\mathcal{H}$, so we would have to invoke the Radon-Nikodym theorem to consistently

§5, one of the two possible theories in $\mathcal{H}_C$ will be much more likely than the other.

The story becomes more complicated if all theories are equally likely (Pr is uniform) or multiple theories are tied as the most likely. In this case, we do not get a unique prediction for the agent's posterior beliefs. The agent could be free to choose any of the most likely hypotheses, or to retain uncertainty over which theory is correct. If the agent retains uncertainty, he or she could keep the entire distribution Pr in mind, yielding an updated belief state with

$$\Omega_C = \bigcup_{h \in \mathcal{H}_C} \Omega_h$$

and probability calculated as:

$$\pi_C(\omega) = \sum_{h \in \mathcal{H}_C} \Pr(h) \times \pi_h(\omega).$$

This new belief state includes all outcomes which are possible in any of the theories and determines the likelihood of each outcome as the expectation of how likely each outcome is, where the expectation is calculated over all possible theories, weighing theories according to Pr.

While the hypothesis testing model does offer determinate predictions for the agent's posterior distribution in many learning situations, the fact that some situations leave the posterior indeterminate represents one way the hypothesis testing model differs from the information minimization model.[7] When there are too many plausible alternative hypotheses consistent with the new constraint, there is no rationally required posterior distribution and multiple solutions to the learning problem can be justified. In §6, I argue that this is the case with the Judy Benjamin problem: there are too many possible hypotheses which explain the new constraint for the hypothesis testing model to offer robust predictions, consistent with the disagreement about how to solve the problem in the literature.

We can summarize the hypothesis testing model for learning a theory-level constraint C in the following steps:

1. The agent has a theory $h = (\Omega, \pi)$ and a constraint C which is incompatible with the theory h.

2. The agent constructs a space $\mathcal{H}$ of alternative hypotheses to h and determines $\mathcal{H}_C$, the subset of hypotheses consistent with C. The agent determines a prior Pr over $\mathcal{H}_C$ representing how likely each theory is.

3. If there is a most likely theory under Pr, the agent chooses that theory. If not, the agent either chooses a theory at random from the most likely theories or retains the entire distribution over theories, keeping a two-level probability space in mind.

apply Bayesian updating. See the discussion of conditional expectations in Ch. 7.4.1 of Capinski and Kopp (2013).

[7] Note that one can incorporate information minimization in the hypothesis testing model. One could argue that, for all learning problems, the hypothesis space consists of all possible probability distributions over the sample space and that the distribution Pr over $\mathcal{H}_C$ tracks the relative information between a new hypothesis h' and the original hypothesis h. In this case, the highest probability alternative theory will be the one which minimizes the distance from the prior, so the predictions of the hypothesis testing model and minimum information model coincide. As argued in this paper, however, this is often not the most natural choice for the hypothesis space.

This procedure will become clearer when we consider some examples: probabilistic learning in §4, conditional learning in §5, and learning conditional probability constraints in §6.

4 Learning Probabilistic Information

Given a belief state (Ω, π) and proposition $A \subseteq \Omega$, a probabilistic constraint is a constraint of the form $\Pr(A) = p$, or in words, 'The probability that A is p.' To consider one case where hypothesis testing provides a useful approach to learning probabilistic information, we consider a simple coin flipping example. Suppose someone is flipping a coin twice, so the set of possible outcomes is $\Omega = \{HH, HT, TH, TT\}$, where H represents heads and T represents tails. The agent assumes the coin is fair, so for each state ω, $\pi(\omega) = \frac{1}{4}$.

Now suppose the agent learns the probabilistic constraint 'The probability the first coin flip is heads is 0.7,' or $\Pr(HH \vee HT) = 0.7$. This constraint is not satisfied on the initial theory where the coin is fair: for (Ω, π), $\pi(HH \vee HT) = 0.5$. Furthermore, there is no event corresponding to the probabilistic constraint and there is no straightforward way to change the initial theory so that this constraint is satisfied. Note how this contrasts with the case where one learns factual or observational information: if one learns that the first coin flip landed heads, it is straightforward to incorporate this information through Bayesian learning by eliminating TH and TT from the set of possible outcomes.

The hypothesis testing model proposes that the agent incorporates this constraint by considering a salient space of alternative hypotheses to the theory that the coin is fair. A natural approach is for the agent to assume that the coin is biased and that the probabilistic constraint is communicating information about the coin bias. In this case, the agent is using a parametric model, where the parameter governing the hypothesis space is the coin bias $\theta \in [0, 1]$, where the coin lands heads with probability θ and tails with probability $1 - \theta$. Note that this hypothesis space is a restriction of the space of all possible theories, excluding, for example, theories where the coin bias varies over time, and that the choice of this hypothesis space is what offers a prediction for how the agent updates beliefs.

The initial theory is that the coin is unbiased, so $\theta = 0.5$. When this theory proves inconsistent with the new probabilistic information, the agent considers other possible biases the coin may have. Intuitively, we expect the agent to update beliefs so that the bias of the coin is now 0.7, since this is the only theory consistent with the probability of the first coin flip landing heads being 0.7.

In fact, we can show that $\theta = 0.7$ is the only hypothesis which satisfies the probabilistic constraint, so that if $\mathcal{H} = [0, 1]$, $\mathcal{H}_C = \{0.7\}$. We can show this by calculating, for every bias θ, the full distribution over outcomes, π_θ: $\pi_\theta(HH) = \theta^2$, $\pi_\theta(HT) = \pi_\theta(TH) = \theta(1 - \theta)$, and $\pi_\theta(TT) = (1 - \theta)^2$. Recall that we formalized a theory as a probability distribution over Ω: here, the parameter θ summarizes the information of a complete theory (Ω, π_θ). A hypothesis θ satisfies C if $\Pr(HH \vee HT) = 0.7$, which occurs when $\theta^2 + \theta(1 - \theta) = 0.7$; $\theta = 0.7$ is the only hypothesis satisfying this criterion. Since $\mathcal{H}_C = \{0.7\}$ has only one element, the agent only has one hypothesis to choose from which satisfies the constraint, so the agent would adopt this theory. On this theory, where the coin bias is $\theta = 0.7$, the new distribution over Ω after learning the constraint C, π_C, is given by $\pi_C(HH) = 0.49$, $\pi_C(HT) = \pi_C(TH) = 0.21$, and $\pi(TT) = .09$.

We can contrast this with the predictions offered by Jeffrey conditioning, which is the approach to incorporating probabilistic constraints which minimizes the Kullbeck-Leibler divergence (Jeffrey, 1990). On Jeffrey conditioning, given a constraint $\Pr(A) = p$, a distribution π is updated to π_J by $\pi_J(\omega) = p\pi(\omega|A) + (1-p)\pi(\omega|\neg A)$.[8] In the above case where $A = HH \vee HT$ and $p = 0.7$, we get the following updated distribution: $\pi_J(HH) = \pi_J(HT) = 0.35$ and $\pi_J(TH) = \pi_J(TT) = 0.15$. This means that the probability that the second coin flip will land heads is 0.5. This follows from the fact that the updated distribution minimizes the information changed from the prior distribution: the new information changes the expectation for the first coin flip, but not the second coin flip. However, it is more intuitive to suppose that the information which changes the bias of the first toss carries over to the second coin toss, showing how theory-based learning offers a more plausible learning prediction than the relative information minimizing approach.

5 Conditional Learning

Another class of constraints which can trigger hypothesis revision are conditional constraints. Assume we have a theory (Ω, π) where Ω represents the set of possibilities and π determines the relative likelihood of these possibilities. Suppose $A, B \subseteq \Omega$ are two propositions and consider the conditional proposition 'If A, then B,' represented $A \to B$. We can think of $A \to B$ as being true relative to the theory (Ω, π) if all A-worlds are B-worlds, so $A \subseteq B$. This follows the standard quantificational approach to the conditional (Lewis, 2013; Kratzer, 1986; Gillies, 2010).

To show how theory revision plays a role in conditional learning, we consider an example where the information minimization method fails: the sundowners problem (Douven and Romeijn, 2011; Douven, 2012). In this case, we suppose that someone considers the possibility that it will rain later (R) and predicts whether sundowners (a kind of party) will occur (S). We assume the agent initially thinks these events are independent and each has probability $\frac{1}{2}$. We can represent the original theory by a set of possibilities $\Omega = \{R \wedge S, R \wedge \neg S, \neg R \wedge S, \neg R \wedge \neg S\}$, each with probability $\frac{1}{4}$. Suppose this person then learns the constraint 'If it rains, then sundowners will be canceled,' $R \to \neg S$.

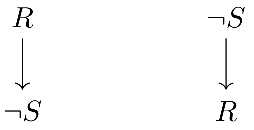
Since the agent represents it as a relevant possibility with probability $\frac{1}{4}$ that it rains and sundowners occurs, $R \wedge S$, this conditional is false on the agent's initial model. This means that the agent must find a new theory which, at minimum, eliminates $R \wedge S$ as a possibility since the new conditional constraint rules it out. The new theory which satisfies this constraint while changing as little as possible is the theory where one eliminates $R \wedge S$ from Ω and redistributes π according to Bayesian updating, so the other three possibilities all have probability $\frac{1}{3}$. This is the approach which minimizes the Kullback-Leibler divergence and is equivalent to updating by the material conditional, where we only eliminate the worlds where $R \supset \neg S$ is false, which are the worlds where R is true and S is true, $R \wedge S$. This updating procedure, however, has counterintuitive results. In particular, in the new distribution, $\Pr(R) = \frac{1}{3}$, meaning that the agent now thinks it is less likely to rain than he or she did initially. However, the conditional constraint that sundowners will be canceled in the event of rain does not communicate any

[8] Note that a choice of partition is required for Jeffrey conditioning (Diaconis and Zabell, 1982). Here, I assume the finest partition into worlds $\omega \in \Omega$. While one could argue that an alternative partition is more suitable, this seems to import the reasoning of hypothesis testing: the set of possible partitions forms a hypothesis space and the agent chooses the most likely hypothesis from this space.

information about whether it will rain, so it seems that the agent should keep $\Pr(R)$ constant at $\frac{1}{2}$.

Assuming that the conditional requires developing a new theory of the world provides more intuitive results. Initially, the agent believes that rain and sundowners are independent of each other, so that neither variable has any effect on (or is relevant for determining) the other variable. However, the conditional can only be true if there is some relationship between the variables, so the agent must consider different possible theories about how the variables are related. The relationships between variables can be encoded in causal models, so the space of alternative hypotheses the agent considers is a set of causal models.

In this example, since R and S are the only variables at issue, we consider only two simple causal models which could satisfy the conditional constraint. We represent these models graphically, where the nodes refer to variables and the arrows represent causal influence:



The nodes at the top of the diagrams without parents are the independent variables, and the nodes below them are dependent variables whose values are determined by the parents.[9]

These diagrams only encode how the variables are related to each other, and many probability distributions are consistent with these causal diagrams. We consider only elements of $\mathcal{H}_C$, or probability distributions over the above causal diagrams which satisfy the conditional constraint. While many such distributions are possible, Vandenburgh (2020) argues that there is a natural way to use a conditional constraint and causal diagram to uniquely determine a new distribution. This procedure assumes that the probability of an independent variable remains constant, but that the probability of the dependent variable changes based on the conditional information: for example, in the first diagram where R causes $\neg S$, while $\Pr(R)$ and $\Pr(\neg R)$ remain the same, $\Pr(\neg S)$ changes according to the formula $\Pr(\neg S) = \Pr(\neg S|R)\Pr(R) + \Pr(\neg S|\neg R)\Pr(\neg R)$. The unique new probability distribution associated with each causal diagram is given below; a more complete justification for this procedure can be found in Vandenburgh (2020).

On the first theory, rain is an independent variable, so $\Pr(R)$ stays constant at $\frac{1}{2}$. Since R causes $\neg S$, the conditional constraint specifies that this causal connection is universal, so whenever R is true, $\neg S$ is true. This means that $R \wedge S$ is not a possibility and that $\Pr(\neg S|R) = 1$, so $\Pr(R \wedge \neg S) = \frac{1}{2}$. Since the conditional constraint gives us no information about the likelihood of S without rain, we assume $\Pr(S|\neg R) = \frac{1}{2}$ as before, so $\Pr(\neg R \wedge S) = \Pr(\neg R \wedge \neg S) = \frac{1}{4}$.

On the second theory, $\neg S$ has a causal influence on R, so S is the independent variable, meaning $\Pr(S) = \Pr(\neg S) = \frac{1}{2}$. The conditional constraint is incorporated through the requirement that $\neg S$ is the only cause of R, so whenever R is true, we can deduce that $\neg S$ is true. This means that $R \wedge S$ is not a possibility and $\Pr(\neg R \wedge S) = \frac{1}{2}$. When $\neg S$ is true, we assume the likelihood of R is 50%, since even though we have learned that $\neg S$ has a causal influence on R, we have not learned anything about how strong this causal influence is, so we retain the prior beliefs. This means that $\Pr(R \wedge \neg S) = \Pr(\neg R \wedge \neg S) = \frac{1}{4}$.

[9] For a more detailed introduction to causal models, see Pearl (2009).

Since both theories satisfy the conditional constraint C, they are both are in the hypothesis space $\mathcal{H}_C$. However, the first theory is much more reasonable than the second theory. It is likely that rain causes the cancelation of a party, but it is very unlikely that the fact that a party is canceled could cause it to rain: it is known that decisions about events do not have an effect on the weather. This means that the agent's probability distribution Pr on the hypothesis space renders the first theory significantly more likely than the second theory, so that the new theory the agent adopts is the first theory. Notice that this theory offers very reasonable predictions: the probability of rain remains constant, the probability of sundowners occurring decreases, and the agent now attributes a causal connection between rain and the cancelation of sundowners. These predictions are much more plausible than those offered by the information minimization theory, which predicts that the likelihood of rain decreases.

6 Conditional Probability Constraints: Judy Benjamin

Information minimization methods of learning have perhaps been most controversial in the case of conditional probability constraints. When A and B are propositions defined over a set of worlds Ω, a conditional probability constraint is a constraint of the form $\Pr(B|A) = p$, with $p \neq 0, 1$. The problem which has drawn attention to the limitations of information minimization is the Judy Benjamin problem (Van Fraassen, 1981). This problem is based on the movie *Private Benjamin*, where Judy Benjamin is dropped into a swampy area and is trying to figure out where she is. The area is divided into two parts, enemy (Red) territory and friendly (Blue) territory, and these areas are both subdivided into Headquarters Company Area (1) and Second Company Area (2). We assume each of the four divisions have equal area, representing the map as follows:

R2	R1
B2	B1

Here, $R1$ and $B1$ represent Red and Blue Headquarters Areas and $R2$ and $B2$ represent Red and Blue Second Company Areas.

Benjamin's radio is largely ineffective, but she receives the information 'If you are in enemy territory, the probability you are in HQ is $\frac{3}{4}$.' This is a conditional probability constraint, where if R represents the Red territory, the constraint says $\Pr(R1|R) = \frac{3}{4}$. Since Benjamin is initially uncertain about where she was dropped and each section has equal area, her prior beliefs ascribe equal probability $\frac{1}{4}$ to each possibility in $\Omega = \{R1, R2, B1, B2\}$. Under these prior beliefs, the conditional probability constraint is not met since $\pi(R1|R) = \frac{1}{2}$, so she must revise her beliefs in line with the new information.

Van Fraassen (1981) works out how Benjamin should change her beliefs in light of this conditional probability constraint if she were to minimize the Kullback-Leibler divergence between her new beliefs and her prior beliefs. In this case, Benjamin ends up decreasing her estimation of the likelihood that she is in enemy territory. This is analogous to the problem identified for conditionals in the previous section, where learning a conditional constraint by minimizing information always leads to a lower probability for the antecedent of the conditional.[10] van Fraassen concludes that this updating procedure is more likely to be the result of wishful thinking than rational updating, since Benjamin concluded that the odds she is in enemy territory are lower from a constraint which appears to say nothing about how likely it is that she is in enemy territory.

We can reconsider this problem from the perspective of hypothesis revision. Initially, Benjamin is completely uncertain about where she has landed on the map; this is consistent with the theory that she was dropped randomly on the map, leading to a uniform distribution over the four quadrants. When Benjamin learns the conditional information that $\Pr(R1|R) = \frac{3}{4}$, she must realize that this is inconsistent with her hypothesis. If she were dropped at random over the area of the map, this conditional probability constraint would not hold, so her hypothesis must be wrong. This leads her to consider alternative hypotheses which are consistent with the new constraint. If the new credences are the result of hypothesis revision, we see that the problem is underdetermined by the information given in the problem: there are multiple ways Benjamin could revise her hypothesis which are consistent with the new information.[11]

For illustration, here are three approaches Benjamin could take to restrict a hypothesis space to take into account the conditional probability constraint. (1) Benjamin could infer that the constraint arises because there is a small off-limits area that she could not be dropped in. Given the fact that the constraint says she is more likely to be in $R1$ than $R2$, she would likely infer that the off-limits area is in $R2$, making it less likely she is in enemy territory. (2) Benjamin could hypothesize about where the intended target was when she was dropped, concluding that the constraint arises because the intended target was either in $R1$ or $B1$. In this case, the likelihood of her remaining in enemy territory remains the same. (3) Benjamin could infer that the constraint arises because of knowledge about the wind conditions. The increased likelihood of her landing in $R1$ compared to $R2$ likely results from the fact that the wind is blowing to the right, which again leaves the likelihood of landing in enemy territory unchanged.

To illustrate how the hypothesis revision approach works formally in this case, I consider (2) in greater detail, where Benjamin forms a new theory about where the intended target was when she was dropped. For the probabilistic constraint to be met, we must assume that the target is not met with perfect accuracy. Thus, if the target is quadrant Q, we can assume that the probability of landing in Q is $\frac{9}{16}$, the probability of landing in one of the two neighboring quadrants is $\frac{3}{16}$, and the probability of landing in the diagonal quadrant is $\frac{1}{16}$. This means that Benjamin considers four theories in $\mathcal{H}$: that the target was $R1$, $R2$, $B1$, or $B2$. Calculating the full distributions, the only two targets which would give rise to the constraint $\Pr(R1|R) = \frac{3}{4}$ are $R1$ and $B1$; thus, these are the only two hypotheses in $\mathcal{H}_C$.

Absent further information, Benjamin would assign equal probability $\frac{1}{2}$ to each hypothesis, but other background information (such as who dropped her and for what purpose) could weigh in favor of one hypothesis over the other. Assuming Benjamin has no fur-

[10] See the discussion in Douven and Romeijn (2011) and Gaifman and Vasudevan (2012).

[11] This is also the conclusion of Bovens and Ferreira (2010), who use Shafer's (1985) theory of protocols to argue that the 'structure of the game' is underdetermined.

ther information, she has two hypotheses which are equally likely, so it makes sense for her to keep both in mind. Keeping both hypotheses in mind, her new beliefs are given by $\pi_C(Q) = \frac{1}{2}\pi_{R1}(Q) + \frac{1}{2}\pi_{B1}(Q)$. For example, the new likelihood that she is in enemy territory is $\pi_C(R) = \frac{1}{2}(\frac{3}{16} + \frac{9}{16}) + \frac{1}{2}(\frac{1}{16} + \frac{3}{16}) = \frac{1}{2}$, showing that the likelihood of her being in enemy territory remains the same.

This illustrates one way in which hypothesis learning can be applied to the Judy Benjamin problem. Benjamin realizes her initial hypothesis, that she was dropped at random, cannot account for the new information, so she must revise this to some other hypothesis which fits the data. Here, her new hypothesis is that whoever dropped her intended for her to land in either $R1$ or $B1$. However, she could have considered a number of other ways to form a new hypothesis depending on how she believes the information given to her was determined, suggesting that the problem as traditionally posed is underdetermined by the information provided.

7 Conclusion

This paper argues that the process of learning new information is less conservative than typically assumed in models of belief revision. While Bayesian learning offers a powerful framework for simple constraints, it fails to generalize to complex, theory-laden propositions like conditionals and probabilistic sentences. Furthermore, conservative extensions of the Bayesian model, like relative information minimization, yield counterintuitive predictions in many cases. In contrast, the model of hypothesis testing, where agents engage in sometimes radical revision of their prior theories, offers more promising results for learning complex constraints. In this model, agents start with a theory of the world, represented by a probability distribution over outcomes (Ω, π), which proves inconsistent with some constraint C. Upon learning C, agents choose the most likely hypothesis h' from the set of hypotheses satisfying C, $\mathcal{H}_C$. The main feature of the hypothesis testing model which offers predictions for learning is the determination of the hypothesis space $\mathcal{H}$, which the agent develops from the context of the learning problem using tools like parametrization and causal models. This provides reasonable predictions for specific examples of learning, such as probabilistic constraints ('The probability the first coin flip is heads is 0.7') and conditional constraints ('If it rains, then sundowners will be canceled'). Furthermore, it explains how the problem of learning can be underdetermined, as I argue is the case in the Judy Benjamin problem. While I argue the hypothesis testing model is successful in these cases where information minimization fails, more research is needed to understand how the model generalizes and whether its predictions can be vindicated through empirical work.

References

Alchourrón, C. E., Gärdenfors, P. and Makinson, D. (1985), 'On the logic of theory change: Partial meet contraction and revision functions', *Journal of symbolic logic* pp. 510–530.

Bovens, L. and Ferreira, J. L. (2010), 'Monty Hall drives a wedge between Judy Benjamin and the Sleeping Beauty: a reply to Bovens', *Analysis* **70**(3), 473–481.

Capinski, M. and Kopp, P. E. (2013), *Measure, integral and probability*, Springer Science & Business Media.

Carey, S. (2009), *The origin of concepts*, Oxford University Press.

Carey, S. and Spelke, E. (1994), 'Domain-specific knowledge and conceptual change', *Mapping the mind: Domain specificity in cognition and culture* **169**, 200.

Casella, G. and Berger, R. L. (2002), *Statistical inference*, Vol. 2, Duxbury Pacific Grove, CA.

Diaconis, P. and Zabell, S. L. (1982), 'Updating subjective probability', *Journal of the American Statistical Association* **77**(380), 822–830.

Douven, I. (2012), 'Learning conditional information', *Mind & Language* **27**(3), 239–263.

Douven, I. and Romeijn, J.-W. (2011), 'A new resolution of the Judy Benjamin problem', *Mind* **120**(479), 637–670.

Eva, B., Hartmann, S. and Rad, S. R. (2020), 'Learning from conditionals', *Mind* **129**(514), 461–508.

Foster, D. P. and Young, H. P. (2003), 'Learning, hypothesis testing, and Nash equilibrium', *Games and Economic Behavior* **45**(1), 73–96.

Gaifman, H. and Vasudevan, A. (2012), 'Deceptive updating and minimal information methods', *Synthese* **187**(1), 147–178.

Gärdenfors, P. (1988), *Knowledge in flux: Modeling the dynamics of epistemic states.*, The MIT press.

Gillies, A. S. (2006), 'What might be the case after a change in view', *Journal of Philosophical Logic* **35**(2), 117–145.

Gillies, A. S. (2010), 'Iffiness', *Semantics and Pragmatics* **3**, 1–42.

Goldstein, S. (2020), 'Epistemic modal credence', *Philosophers' Imprint* pp. 1–34.

Gopnik, A. (1996), 'The scientist as child', *Philosophy of science* **63**(4), 485–514.

Griffiths, T. L. and Tenenbaum, J. B. (2009), 'Theory-based causal induction', *Psychological review* **116**(4), 661–716.

Halpern, J. Y. (2017), *Reasoning about uncertainty*, MIT press.

Jeffrey, R. C. (1990), *The logic of decision*, University of Chicago Press.

Kratzer, A. (1986), 'Conditionals', *Chicago Linguistics Society* **22**(2), 1–15.

Kratzer, A. (2012), *Modals and Conditionals: New and Revised Perspectives*, Vol. 36, Oxford University Press.

Kuhn, T. S. (2012), *The structure of scientific revolutions*, University of Chicago press.

Lewis, D. (2013), *Counterfactuals*, John Wiley & Sons.

Ortoleva, P. (2012), 'Modeling the change of paradigm: Non-Bayesian reactions to unexpected news', *American Economic Review* **102**(6), 2410–36.

Pearl, J. (2009), *Causality*, Cambridge university press.

Rehder, B. (2003), 'A causal-model theory of conceptual representation and categorization.', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **29**(6), 1141.

Salmon, T. C. (2004), 'Evidence for learning to learn behavior in normal form games', *Theory and decision* **56**(4), 367–404.

Shafer, G. (1985), 'Conditional probability', *International Statistical Review/Revue Internationale de Statistique* pp. 261–275.

Van Fraassen, B. C. (1981), 'A problem for relative information minimizers in probability kinematics', *The British Journal for the Philosophy of Science* **32**(4), 375–379.

Vandenburgh, J. (2020), 'Conditional learning through causal models', *Synthese* . https://doi.org/10.1007/s11229-020-02891-x.

Veltman, F. (1996), 'Defaults in update semantics', *Journal of philosophical logic* **25**(3), 221–261.

Wellman, H. M. and Gelman, S. A. (1992), ‘Cognitive development: Foundational theories of core domains’, *Annual review of psychology* **43**(1), 337–375.

Williams, P. M. (1980), ‘Bayesian conditionalisation and the principle of minimum information’, *The British Journal for the Philosophy of Science* **31**(2), 131–144.

Yalcin, S. (2010), ‘Probability operators’, *Philosophy Compass* **5**(11), 916–937.

Yalcin, S. (2012), Context probabilism, *in* ‘Logic, language and meaning’, Springer, pp. 12–21.

Young, H. P. (2004), *Strategic learning and its limits*, OUP Oxford.