Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

Dynamics of Inductive Inference in a Unified Framework[1]

Itzhak Gilboa,[2] Larry Samuelson,[3] David Schmeidler[4]

May 22, 2012

Abstract

We present a model of inductive inference that includes, as special cases, Bayesian reasoning, case-based reasoning, and rule-based reasoning. This unified framework allows us to examine how the various modes of inductive inference can be combined and how their relative weights change endogenously. For example, we establish conditions under which an agent who does not know the structure of the data generating process will decrease, over the course of her reasoning, the weight of credence put on Bayesian vs. non-Bayesian reasoning. We illustrate circumstances under which probabilistic models are used until an unexpected outcome occurs, whereupon the agent resorts to more basic reasoning techniques, such as case-based and rule-based reasoning, until enough data are gathered to formulate a new probabilistic model.

[1]We thank Daron Acemoglu, Dirk Bergemann, Eddie Dekel, Drew Fudernberg, Gabi Gayer, Offer Lieberman, George Mailath, an associate editor, and two referees for comments and suggestions.

[2]Tel-Aviv University, HEC, Paris, and Cowles Foundation, Yale University. ISF Grant 396/10 and ERC Grant 269754 are gratefully acknowledged.

[3]Department of Economics, Yale University. National Science Foundation grant SES-0850263 is gratefully acknowledged.

[4]Tel-Aviv University, The Ohio State University, and the InterDisciplinary Center in Herzliya. ISF Grant 396/10 is gratefully acknowledged.

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

May 22, 2012

Contents

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

May 22, 2012

1 Introduction

Economic theory typically assumes that agents reason about uncertainty in a Bayesian way: they formulate prior probabilities over a state space and update them in response to new information according to Bayes' rule. This model is powerful, but does not always reflect the way that people think about uncertainty. In particular, when completely unexpected outcomes occur, people question their probabilistic models, relying on alternative reasoning techniques until perhaps developing a new probabilistic model.

For example, the New York Stock Exchange was closed for five days following the September 11, 2001 terrorist attacks on the United States. On the following Sunday, September 16, a leading economist was asked to predict the behavior of the Dow Jones Industrial Average on Monday. He did not respond by reasoning that "I used to attach (the quite small) probability ε to such attacks, and now I need only update this probability, and then apply my usual model of the stock market." Instead, there was a sense that the probabilistic model he would have used under normal circumstances was inappropriate for the present situation, and that he had to start from basics in reasoning about the future. He responded by invoking analogies to past cases in which the United States had been surprised by attack, most notably Pearl Harbor. (As it turned out, his prediction was quite accurate.)

Similarly, following the collapse of Lehman Brothers in September 2008, the head of a major investment firm confronted clients anxious to sell their assets, even assets that had already lost 90% of their value. Again, the analyst did not apply Bayes' rule to a prior that had taken into account a possible failure of Lehman Brothers. Instead, he argued that something totally unexpected had happened, and that "obviously, the models do not work." The analyst convinced his clients to hold such assets, invoking the simple rule that "an asset that has lost 90% of its value cannot lose much more." (His clients were convinced, and subsequently appreciated the advice.)

In both examples, one could, post-hoc, construct a prior probability distribution that allows the experts' reasoning to follow from Bayesian updating. However, such a description would say very little about the actual reasoning

process of the agents involved, and (more importantly) would not be of much help in predicting their reasoning in the future. Our interest in this paper is in modeling the agents' actual reasoning processes, in the hope of better understanding when these processes generate probabilistic beliefs, which beliefs are likely to be formed by the agents, and how the agent might form beliefs when driven away from familiar probabilistic models. To do so, we need a model that can simultaneously describe probabilistic and non-probabilistic reasoning, as well as the dynamics by which weights shift between modes of reasoning.

We take it for granted that when statistical analysis is possible, rational agents will perform such analysis correctly. In contrast, our interest is in the way economists model agents who face problems that do not naturally lend themselves to statistical analysis. Predicting financial crises, economic growth, the outcome of elections, or the eruptions of wars and revolutions, are examples where it is difficult to define iid random variables and, more generally, where the assumptions of statistical models do not seem to be good approximations.

To set the context for our model, consider an agent who each year is called upon to predict the price of oil over the subsequent year. To keep this illustrating example simple, suppose the agent need only predict whether the average price will be higher or lower than the previous year's price. We can imagine the agent working for a hedge fund that is interested in whether it should bet for or against an increasing price.

To support her decision, the agent's research staff regularly compiles a list of data potentially relevant to the price of oil, as well as data identifying past values of the relevant variables and past oil prices. For our example, let us assume that the data include just two variables: a measure of the change in the demand for oil and a measure of the change in the severity of conflict in the Middle East. Each is assumed to take two values, indicating whether there has been an increase or decrease. Each year the agent receives the current changes in demand and in conflict, examines the data from previous years, and then predicts whether the price will increase or decrease. How do and how should agents reason about such problems?

Our model captures three types of reasoning.[1] The most common in

[1]In personal conversation, a hedge fund principal indicated that his fund used all three methods of reasoning introduced in this section in predicting the likelihood of mortgage defaults.

economic modeling is *Bayesian*. The agent first formulates the set of possible states of the world, where a state identifies the strength of demand, the measure of conflict, and the price of oil, in each year over the course of her horizon. The agent then formulates a prior probability distribution over this state space. This prior distribution will reflect models and theories of the market for oil that the agent finds helpful, her analysis of past data and past outcomes in this market, and any other prior information she has at her command. Once this prior has been formulated, the agent's predictions are a relatively straightforward matter of applying Bayes' rule, as new observations allow her to rule out some states and condition her probability distribution on the surviving states.

An alternative mode of reasoning is *case-based*. The agent considers past observations and predicts the outcome that appeared more often in those past cases that are considered similar. For example, predicting that following the September 11 attacks, the DJIA would change in a similar way to its change following Pearl Harbor would be considered case-based reasoning. If all past observations are considered equally similar, the case-based prediction is simply the mode, that is, the outcome that is most frequent in the database. If the agent uses a similarity function that puts all its weight on the most recent outcome, her prediction will simply be that outcome.[2] If the agent views the current state of conflict in the Middle East as a repetition of affairs in 1991 or in 2003, the agent will predict that there will soon be a war and an increase in the price of oil.

Finally, *rule-based* reasoning calls for the agent to base her predictions on regularities that she believes characterize the market for oil. For example, the agent may adopt a rule that any increase in the level of demand leads to an increase in the price of oil. Based on this and her expectation that the Chinese economy will continue to grow, the agent might reasonably predict that the price is about to rise.

The boundaries between the three modes of reasoning are not always sharp. Our focus is on the Bayesian approach. By "Bayesian reasoning" we refer to the common approach in economic theory, according to which *all* reasoning is Bayesian. *Any* source of uncertainty is modeled in the state space, and all reasoning about uncertainty takes the form of updating a prior probability via Bayes' rule.

[2]Indeed, Alquist and Kilian [2] find that the best prediction of the future price of oil is the current price.

This paper presents (in Sections 2–3) a framework that unifies these three modes of reasoning (and potentially others), allowing us to view them as special cases of a general learning process. The agent attaches weights to conjectures. Each conjecture is a set of states of the world, capturing a way of thinking about how outcomes in the world will develop. The associated weights capture the relative influence that the agent attaches to the various conjectures. To generate a prediction, the agent sums the weight of all nontrivial conjectures consistent with each possible outcome, and then ranks outcomes according to their associated total weights. In the special case where each conjecture consists of a single state of the world, our framework is the standard Bayesian model, and the learning algorithm is equivalent to Bayesian updating. Employing other conjectures, which include more than a single state each, we can capture other modes of reasoning, as illustrated by simple examples of case-based and of rule-based reasoning.

Our model could be used to address either positive or normative questions. In this paper we focus on positive ones, describing how the reasoning process of an agent evolves as observations are gathered. Within the class of such questions, our model could be used to capture a variety of psychological biases and errors, but the focus of this paper is on the reasoning of an agent who makes no obvious errors in her reasoning. Such an agent may well be surprised by circumstances that she has deemed unlikely, that is, by "black swans," but will never be surprised by a careful analysis of her own reasoning. The optimality of this reasoning process is a normative question, which we do not address here.

Our main results concern the dynamics of the weight put on Bayesian vs. non-Bayesian reasoning. In Section 4.1 we suggest conditions under which Bayesian reasoning will give way to other modes of reasoning, and alternative conditions under which the opposite conclusion holds. Section 4.3 briefly discusses how probabilistic reasoning may emerge periodically, with other modes of reasoning used between the regimes of different probabilistic models. Section 5 concludes.

2 The Framework

2.1 The Environment

At each period $t \in \{0, 1, \ldots\}$ there is a *characteristic* $x_t \in X$ and an *outcome* $y_t \in Y$. The sets X and Y are assumed to be finite and non-empty, with Y containing at least two possible outcomes.[3]

In predicting the price of oil, the characteristic x_t might identify the type of political regime and the state of political unrest in various oil-producing countries, describe the extent of armed conflict in the Middle East, indicate whether new nuclear power plants have come on line or existing ones been disabled by accidents, describe the economic conditions of the major oil importers, summarize climate conditions, and so on. In our simplified example, Y has only two elements, $\{0, 1\}$, and each $x = (x^1, x^2) \in X$ has two components, each also taking values in $\{0, 1\}$, with a 1 in each case indicating an increase in the relevant variable.

We make no assumptions about independence or conditional independence of the variables across periods. In fact, for most of our analysis we do not assume any probability on the state space, so that independence of the variables cannot even be defined. The model can be augmented by assumptions about the underlying probability measure that drives the process, allowing one to state results about learning the "true" state of the world. While some of our examples below are of this nature, the general framework is silent on the actual data generating process.

A *state of the world* ω identifies the characteristic and outcome that appear in each period t, i.e., $\omega : \{0, 1, ...\} \to X \times Y$. We let $(\omega_X(t), \omega_Y(t))$ denote the element of $X \times Y$ appearing in period t given state ω, and let

$$\Omega = (X \times Y)^\infty$$

denote the set of states of the world. In our example, a state identifies the sign of changes in the strength of demand, the level of conflict, and the price of oil in each period.

A period-t history

$$h_t = (\omega_X(0), \omega_Y(0), \ldots, \omega_X(t-1), \omega_Y(t-1), \omega_X(t))$$

identifies the characteristics (e.g., changes in the levels of demand and of conflict) and outcomes (e.g., changes in the price of oil) that have appeared

[3]The extension to infinite sets X and Y can be carried out with no major difficulties.

in periods 0 through $t-1$, as well as the period-t characteristic, given state ω. We let H_t denote all possible histories at period t. For a particular history h_t we define the corresponding event

$$[h_t] = \{\omega \in \Omega \mid (\omega(0), \ldots, \omega(t-1), \omega_X(t)) = h_t\}$$

consisting of all states that are compatible with the history h_t. In other words, $[h_t]$ is the set of states whose period-t history matches h_t, with different states in this set corresponding to different possible future developments. We define, for $h_t \in H_t$ and $Y' \subset Y$, the event

$$[h_t, Y'] = \{\omega \in [h_t] \mid \omega_Y(t) \in Y'\}$$

consisting of all states that are compatible with the history h_t and with a period-t outcome in the set Y'.

In each period t the agent observes a history h_t and makes a prediction about the period-t outcome, $\omega_Y(t) \in Y$. A *prediction* is a ranking of subsets in Y given h_t. Hence, for $h_t \in H_t$ there is a binary relation $\succsim_{h_t} \subset 2^Y \times 2^Y$ that ranks subsets of outcomes according to their plausibility.

2.2 Predictions

Predictions are made with the help of conjectures. Each conjecture is a subset $A \subset \Omega$. A conjecture can represent a specific scenario, that is, a single state of the world, in which case $A = \{\omega\}$. However, conjectures can contain more than one state, and thereby capture rules and analogies. In general, any reasoning aid one may employ in predicting y_t can be described by the set of states that are compatible with it.

In principle, a conjecture could be *any* subset of Ω, but the set of all subsets of Ω is rather large and unwieldy. Nothing is lost by taking the set of conjectures to be the σ-algebra $\mathcal{A}$ generated by the events $\{[h_t]\}_{t\geq 0, h_t \in H_t}$.[4]

To make predictions in period t, the agent first identifies, for any subset of outcomes $Y' \subset Y$, the set of conjectures that have not been refuted by the history h_t and that predict an outcome in Y'. A conjecture $A \in \mathcal{A}$ has not been refuted by history h_t if $A \cap [h_t] \neq \varnothing$. The set of conjectures that have

[4]Note that this is the same σ-algebra generated by $\{[h_t, Y']\}_{t\geq 0, h_t \in H_t, Y' \subset Y}$ and that it contains all singletons, i.e., $\{\omega\} \in \mathcal{A}$ for every $\omega \in \Omega$.

not been refuted by history h_t *and* predict an outcome in Y' is[5]

$$\mathcal{A}(h_t, Y') = \{A \in \mathcal{A} \mid \varnothing \neq A \cap [h_t] \subset [h_t, Y']\}. \tag{1}$$

The agent evaluates the relative likelihoods of outcomes Y' and Y'', at history h_t, by comparing the sets $\mathcal{A}(h_t, Y')$ and $\mathcal{A}(h_t, Y'')$. The agent makes this comparison by using a *credence function* φ_{h_t}. Formally, φ_{h_t} is a finite, nonzero σ-additive measure on a sigma-algebra $\mathcal{E} \subset 2^{\mathcal{A}}$ to be defined shortly.[6] We interpret $\varphi_{h_t}(\mathcal{A}(h_t, Y'))$ as the weight the agent attaches to conjectures consistent with the outcomes Y', and $\varphi_{h_t}(\mathcal{A}(h_t, Y''))$ as the weight the agent attaches to conjectures consistent with the outcomes Y''.[7]

To make predictions, the agent ranks Y' as "at least as likely as" Y'', denoted $Y' \succsim_{h_t} Y''$, iff

$$\varphi_{h_t}(\mathcal{A}(h_t, Y')) \geq \varphi_{h_t}(\mathcal{A}(h_t, Y^{''})). \tag{2}$$

In Appendix B we provide a characterization of binary ("at least as likely as") relations over subsets of Y that can be represented by belief functions, as $\succsim_{h_t}$ is by represented φ_{h_t} in (2).

Intuitively, one may think of each conjecture A as an expert, who argues that the state of the world has to be in the event A. The weight $\varphi_{h_t}(\{A\})$ is a measure of the expert's reliability in the eyes of the agent. The agent listens to the forecasts of all experts and, when comparing two possible predictions Y' and Y'', chooses the prediction that commands higher total support from the experts. When an expert is proven wrong, he is asked to leave the room and his future forecasts are ignored. For example, upon reaching a history h_t at which the input $\omega_X(t)$ indicates that the demand for oil has declined, perhaps because of a worldwide recession, the oil analyst from Section 1 will discard all conjectures pertaining to the price of oil in conditions of currently expanding demand. The set of surviving conjectures $\mathcal{A}(h_t)$ may include a conjecture consistent with a decrease in the price of oil (perhaps because

[5]Observe that the conjectures $\varnothing$ and Ω are never included in $\mathcal{A}(h_t, Y')$ for any $Y' \subsetneq Y$. The impossible conjecture $\varnothing$ is not compatible with any history h_t, whereas the certain conjecture Ω is tautological at every history h_t.

[6]There is no loss of generality in taking φ_{h_t} to be a probability measure, but it economizes on notation to refrain from imposing this normalization. For example, we thereby avoid the need to constantly make special provision for cases in which denominators are zero.

[7]The weighting function φ_{h_t} is equivalent to a belief function in the Dempster-Shafer theory of evidence (Dempster [9], Shafer [35]). See Appendix B for a brief introduction.

the recession has reduced demand) as well as a conjecture consistent with an increase in the price of oil (perhaps because the price increased during a previous recession that the analyst remembers vividly).

To complete this definition, we need to specify the σ-algebra $\mathcal{E} \subset 2^{\mathcal{A}}$ over which the measures φ_{h_t} are defined.[8] For convenience, the domain of the function φ_{h_t} will be the same σ-algebra $\mathcal{E}$ for each history h_t, even though only a subset of conjectures, namely $\cup_{Y' \subsetneq Y} \mathcal{A}(h_t, Y')$, is relevant for prediction at h_t, and the definition of φ_{h_t} outside this set is irrelevant.

First, for each conjecture $A \in \mathcal{A}$, it will be useful to be able to refer to its weight of credence as $\varphi_{h_t}(\{A\})$, which requires that $\{A\}$ be a measurable set. Let $\mathcal{E}_0$ be the σ-algebra generated by all such sets. Next, since predictions will be made by comparing the φ_{h_t} values of subsets of the type $\mathcal{A}(h_t, Y')$, we need to make sure that these are measurable. Let $\mathcal{E}_1$ be the σ-algebra generated by all such sets. Finally, the set of singletons contained in a conjecture will also be of interest, and we let $\mathcal{E}_2$ be the σ-algebra generated by all such sets.[9] Summarizing:

σ-algebra	Generating sets
$\mathcal{E}_0$	$\{A\}$ for $A \in \mathcal{A}$
$\mathcal{E}_1$	$\mathcal{A}(h_t, Y')$ for $t \geq 0, h_t \in H_t, Y' \subset Y$
$\mathcal{E}_2$	$\{\{\omega\} \,\vert \omega \in A\}$ for $A \in \mathcal{A}$

.

We then define $\mathcal{E}$ as the σ-algebra that is generated by $\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2$. A credence function φ_{h_t} is a (σ-additive) measure on $\mathcal{E}$.

Using states of the world to represent possible outcomes is standard in decision theory, as is the summation of a function such as φ_{h_t} to capture beliefs, and the elimination of conjectures that have been proven wrong. The most obvious departure we have taken from the familiar framework of Bayesian updating is to allow conjectures that consist of more than one state.[10] To confirm this, Section 3.1 shows that if we restrict attention to

[8] Recall that a conjecture A is an element of the σ-algebra $\mathcal{A}$ over the set of states Ω. An element of $\mathcal{E}$ is a set of conjectures, and hence is an element of a σ-algebra over the set $2^{\mathcal{A}}$ of sets of states.

[9] The collection $\mathcal{E}_0$ contains every set of the form $\{\omega\}$, but $\{\{\omega\} \,\vert \omega \in A\}$ may be uncountable, and so must be explicitly included in the definition of the σ-algebra $\mathcal{E}$. Doing so ensures, for example, that the set of Bayesian conjectures is measurable.

[10] In the process, the notion of compatibility needs to be adapted: whereas a single state ω is compatible with history h_t if $\omega \in [h_t]$, a (possibly multistate) conjecture A is compatible with history h_t if $A \cap [h_t] \neq \varnothing$.

singe-state conjectures, then we have the familiar framework of Bayesian reasoning. Expanding the framework to encompass multi-state conjectures is necessary if we are to capture case-based and rule-based reasoning (cf. Sections 3.2 and 3.3).

We have restricted attention to deterministic conjectures. One sees this in (1), where conjectures are either clearly compatible or clearly incompatible with a given history. This is obviously restrictive, as we are often interested in drawing inferences about theories that do not make sharp predictions. However, a framework in which the implications of the evidence for various conjectures is dichotomous simplifies the analysis by eliminating assessments as to which theories are more or less likely for a given history, in the process allowing us to focus attention on the resulting induction.

2.3 Updating

How does the agent learn in this model? We have already identified one avenue for learning, namely that refuted conjectures are thereafter excluded from consideration. If this were the only avenue for learning in our model, then the updating would precisely mimic Bayesian updating, and the only generalization from a standard Bayesian model would be the introduction of multi-state conjectures.

Our generalized model allows a second avenue for learning—the credence function φ_{h_t} can vary with the history h_t. Collecting information allows the agent not only to exclude falsified conjectures, but to modify the weights she attaches to her surviving conjectures. This contrasts with Bayesian updating in a standard probability model, where unrefuted states retain their original relative weights, as well as with the notion of a likelihood function, which can only decrease in value as data are gathered.

We can obviously expect φ_{h_t} to vary with h_t if the agent is initially unaware of some conjectures. Such a conjecture will be assigned a zero weight at the outset, but a positive weight at a history h_t that brings the conjecture to mind. For example, it is possible that prior to September 11, 2001 the agent had not imagined that terrorists might fly commercial airliners into buildings. This unawareness is naturally captured by setting $\varphi_{\varnothing}$ of related conjectures to zero. However, given a history h_t that includes this attack, conjectures that involve similar attacks in the future may have a positive weight in φ_{h_t}.

Even without unawareness, φ_{h_t} may depend on the history h_t. The com-

peting conjectures in our model have different domains of application. Some conjectures make predictions at each period, while others only rarely hazard a prediction. Once we reach a history h_t, shouldn't conjectures that have made many correct predictions along the way be upgraded in comparison to those who have hitherto said little or nothing? In effect, shouldn't the value $\varphi_{h_t}(\{A\})$ increase as A passes more prediction tests?

For example, suppose that there are two possible outcomes ($Y = \{0, 1\}$) and that conjecture A makes predictions at each of the periods $t = 0, ..., 100$, while conjecture A' makes a prediction only at $t = 100$. Conjecture A may be a market analyst who arrives at time $t = 100$ having pegged the market correctly in every period, while conjecture A' may be a competing analyst who thus far has said nothing other than "can't tell."[11] It seems that the weight we attach to A at time $t = 100$ should be higher than that of A', even if at the outset the two analysts seemed equally reliable.

Rewarding conjectures (or experts) for passing more prediction tests does not require that φ_{h_t} depend on h_t. Instead, these rewards can be built into a function φ that is independent of h_t. In the example above, at time $t = 0$ the agent already knows that conjecture A' will be irrelevant for the first 100 observations, and will join the game only at period $t = 100$. The agent can then build this comparison into the function $\varphi_\varnothing$, perhaps by assigning weights $\varphi_\varnothing(A) = 100\varphi_\varnothing(A')$, and can then simply use $\varphi_\varnothing$ throughout. Thus, if at time $t = 100$ conjecture A is still in the game, it will have a much higher weight than would A', without any alteration in φ.[12] In effect, if we know that conjecture A' will take no chances until period 100 and so will then be allocated a small weight relative to whatever conjecture has in the meantime passed many prediction tests, we might as well downgrade A' at the beginning.

Consider a somewhat more involved example in which conjecture A again makes predictions at every period, and A' now makes predictions at periods $t = 0$ and $t = 100$, but remains silent in between. We may then want to assign the two conjectures equal weights at time $t = 0$, but adjust $\varphi_{h_{100}}$ in order to give A credit for having made the intervening string of correct predictions, should both still be relevant at time $t = 100$. It seems as if simply adjusting

[11] For example, we could have $A = [\underbrace{1, \ldots, 1}_{101}]$ and $A' = \{\omega \in \Omega : \omega_Y(100) = 1\}$.

[12] Alternatively, if A predicts incorrectly during some of the first 100 periods, it will subsequently be excluded and hence this choice of $\varphi_\emptyset$ will not interfere with further predictions.

$\varphi_\varnothing$ and thereafter holding φ fixed will not accomplish both goals. However, we can indeed incorporate all of these considerations without allowing φ to depend on h_t. The key is to note that the conjectures A and A' can both be relevant at time $t = 100$ only if they make identical predictions at time $t = 0$. But if they make the same prediction at time $t = 0$, only the sum of their weights (and not their relative weighting) has any effect on predictions at $t = 0$. We can thus freely adjust $\varphi_\varnothing(A)$ and $\varphi_\varnothing(A')$ in such a way that would not change predictions until time $t = 0$, but will give A more weight at time $t = 100$.

The more general point is that $\{\varphi_{h_t}\}_{t\geq 0, h_t \in H_t}$ is under-identified by the rankings $\left\{\succsim_{h_t} \subset 2^Y \times 2^Y\right\}_{t\geq 0, h_t \in H_t}$. Many different credence functions $\{\varphi_{h_t}\}_{t\geq 0, h_t \in H_t}$ give rise to the same ranking of subsets (at each and every history). Indeed it turns out that any ranking that can be obtained by a history-dependent $\{\varphi_{h_t}\}_{t\geq 0, h_t \in H_t}$ can also be represented by a history-independent φ:

Proposition 1 *Let $\{\varphi_{h_t}\}_{t\geq 0, h_t \in H_t}$ be a collection of finite measures on $(\mathcal{A}, \mathcal{E})$. Then there exists a measure φ on $(\mathcal{A}, \mathcal{E})$ such that, at each h_t and for every $Y', Y'' \subset Y$,*

$$\varphi(\mathcal{A}(h_t, Y')) \geq \varphi(\mathcal{A}(h_t, Y'')) \quad \Longleftrightarrow \quad \varphi_{h_t}(\mathcal{A}(h_t, Y')) \geq \varphi_{h_t}(\mathcal{A}(h_t, Y'')).$$

It thus sacrifices no generality to work with a function φ that is unchanged as history unfolds. We accordingly hereafter drop the h_t subscript on φ and work with an unchanging φ.

When φ is independent of history, the updating rule inherent in (1)–(2) is equivalent to the Dempster-Shafer (cf. Dempster [9], Shafer [35]) updating of the belief function defined by φ, in face of the evidence $[h_t]$. (See Appendix B for details.) This updating rule has been axiomatized by Gilboa and Schmeidler [13] in the context of Choquet expected utility maximization.[13]

3 Special Cases

The unified framework is sufficiently general as to capture several standard models of inductive reasoning.

[13]The Dempster-Shafer updating rule, explained in Appendix B, is a special case of Dempster's rule of combination. We mention in passing that it does not suffer from common criticisms of the Dempster-Shafer theory, such as those leveled by Voorbraak [39].

3.1 Bayesian Reasoning

We first show that our framework reduces to Bayesian reasoning if one restricts attention to conjectures that consist of one state each.

Bayesian reasoning has been studied in many ways in many fields.[14] The various manifestations of the Bayesian approach differ in several ways, such as the scope of the state space and the degree to which Bayesian beliefs are related to decision making, but they share two common ingredients: (i) uncertainty is always quantified probabilistically; and (ii) when new information is obtained, probabilistic beliefs are updated according to Bayes' rule.

To embed Bayesian reasoning in our framework, define the set of *Bayesian conjectures* to be

$$\mathcal{B} = \{\{\omega\} \mid \omega \in \Omega\} \subset \mathcal{A}. \tag{3}$$

Notice that $\mathcal{B}$ is an element of $\mathcal{E}$. Moreover, for every history h_t, the set of surviving or unfalsified Bayesian conjectures $\mathcal{B}(h_t)$ is given by

$$\mathcal{B}(h_t) = \{\{\omega\} \mid \omega \in [h_t]\},$$

and it is in $\mathcal{E}$ as well.

A credence function φ is Bayesian if only Bayesian hypotheses matter in determining the weights of credence attached to a set of conjectures, i.e., if for any set $E \in \mathcal{E}$, we have

$$\varphi(E) = \varphi(E \cap \mathcal{B}). \tag{4}$$

In particular, among the conjectures contained in $\mathcal{A}$, only those in $\mathcal{B}$ are assigned positive weight by a Bayesian credence function.

We can confirm that our model captures Bayesian reasoning:

[14] Bayesian reasoning appeared explicitly in the writings of Bayes [3], with precursors from the early days of probability theory such as Bernoulli [4]. Beginning with the work of de Finetti and his followers, it has given rise to the Bayesian approach to statistics (see, for example, Lindley [21]). Relying on the axiomatic approach of Ramsey [28], de Finetti [7, 8], and Savage [32], it has grown to become the dominant approach in economic theory and in game theory. The Bayesian approach has also made significant headways in computer science and artificial intelligence, as in the context of Bayesian networks (Pearl [27]). Within the philosophy of science, notable proponents of the Bayesian approach include Carnap [5] and Jeffrey [20].

Lemma 1 *Let p be a probability measure on $(\Omega, \mathcal{A})$. There exists a Bayesian credence function such that for every history h_t, there is a constant $\lambda > 0$ for which, for every $Y' \subset Y$ and h_t with $p(\mathcal{B}(h_t)) > 0$,*

$$p(y_t \in Y' \mid [h_t]) = \lambda \varphi(\mathcal{A}(h_t, Y')).$$

Proof. Let the credence function be given by

$$\varphi(E) = p(E \cap \mathcal{B}).$$

First note that $E \cap \mathcal{B}$ is in $\mathcal{A}$.

Hence, φ attaches to each set of Bayesian hypotheses a weight of credence equal to the prior probability attached to the set, and attaches to a general set of hypotheses a weight of credence equal to that of the Bayesian hypotheses it contains. Then φ is clearly Bayesian. It then remains to identify the normalization appropriate for history h_t:

$$\begin{aligned}
\varphi(\mathcal{A}(h_t, Y')) &= p(\{\omega : \{\omega\} \in \mathcal{A}(h_t, Y')\}) \\
&= p(\mathcal{B}(h_t)) \frac{p(\{\omega : \{\omega\} \in \mathcal{A}(h_t, Y')\})}{p(\mathcal{B}(h_t))} \\
&= p(\mathcal{B}(h_t)) p(y_t \in Y' \mid [h_t]) \\
&:= \lambda^{-1} p(y_t \in Y' \mid [h_t]),
\end{aligned}$$

giving the result. ■

Bayesian reasoning is thus a special case of our framework: every Bayesian belief can be simulated by a model φ, and Bayesian updating is imitated by our process of excluding refuted conjectures. Apart from the normalization step, which guarantees that updated probabilities continue to sum up to 1 as conjectures are deleted but has no effect on relative beliefs, Bayesian updating is nothing more than the exclusion of refuted conjectures from further prediction.

Our model captures Bayesian reasoning via an assumption that only conjectures containing a single state enter the agent's reasoning. An agent whose credence function assigns positive weight to non-Bayesian conjectures (e.g., $\varphi(\{A\}) > 0$ for some $A \in \mathcal{A} \backslash \mathcal{B} > 0$) will not be "Bayesian" by any common definition of the term. For example, suppose that $A = \{\omega_1, \omega_2\}$ and $\varphi(\{A\}) = \delta > 0$. Such an agent can be viewed as arguing, "I think that one

of ω_1 or ω_2 might occur, and I put a weight $\delta > 0$ on this conjecture, but I cannot divide this weight between the two states." Intuitively, this abandons the Bayesian tenet of quantifying all uncertainty in terms of probabilities. Formally, the corresponding rankings of subsets of outcomes, $\succsim_{h_t}$, may fail to satisfy de Finetti's [7, 8] cancellation axiom: it can be the case that, for two sets of outcomes, B, C, $B \succsim_{h_t} C$ but not $B \backslash C \succsim_{h_t} C \backslash B$. In addition, if we use the credence function to make decisions by maximization of the Choquet integral of a utility function, the maximization will fail to satisfy Savage's [32] "sure-thing principle" (axiom P2). As a result, upon adding decisions to our model of beliefs, we would have a converse to Lemma 1: the decision maker will be Bayesian if and *only if* (4) holds. (See Appendix B for a definition of Choquet integration.)

3.2 Case-Based Reasoning

Case-based reasoning is also a special case of our model.[15]

We first introduce a simple model of case-based reasoning in which case-based prediction is equivalent to kernel classification.[16] The agent has a similarity function over the characteristics,

$$s : X \times X \to \mathbb{R}_+,$$

and a memory decay factor $\beta \leq 1$. Given history $h_t = h_t(\omega)$, a set of outcomes $Y' \subsetneq Y$ is assigned the weight

$$S(h_t, Y') = \sum_{y \in Y'} \sum_{i=0}^{t-1} \beta^{t-i} s(\omega_X(i), \omega_X(t)) \mathbf{1}_{\{\omega_Y(i)=y\}},$$

where $\mathbf{1}$ is the indicator function of the subscripted event. Hence, the agent may be described as if she considered past cases in the history h_t, chose all those that resulted in some period i with some outcome $y \in Y'$, and added to the sum $S(h_t, Y')$ the similarity of the respective characteristic $\omega_X(i)$ to the current characteristic $\omega_X(t)$. The resulting sums $S(h_t, Y')$ can then be used to rank sets of possible outcomes Y'.

[15] Analogical reasoning was explicitly discussed by Hume [19], and received attention in the twentieth century in the guise of case-based reasoning (Riesbeck and Schank [30], Schank [33]), leading to the formal models and axiomatizations of Gilboa and Schmeidler [14, 16, 17].

[16] See Akaike [1] and Silverman [37].

If $\beta = 1$ and in addition the similarity function is constant, the resulting number $S(h_t, \{y\})$ is proportional to the relative empirical frequency of y's in the history h_t. If, on the other hand, $\beta \to 0$, the maximizer of $S(h_t, \cdot)$ will be the most recent observation, $\omega_Y(t-1)$. Thus, when the similarity function is constant, case-based reasoning can be viewed as a simultaneous (and smooth) generalization of prediction by empirical frequencies on the one hand, and of prediction by recency on the other hand.

More interesting generalizations are possible when the similarity function isn't constant, and uses the information given in X to make more informed judgments.

The next observation states that the general framework presented in Section 2 can accommodate a special case of case-based reasoning:

Lemma 2 *Let there be given $s : X \times X \to \mathbb{R}_+$ and $\beta \leq 1$. There exists a credence function φ such that, for every history h_t, there is a constant $\lambda > 0$ for which, for every $y \in Y$,*

$$\varphi(\mathcal{A}(h_t, \{y\})) = \lambda S(h_t, y).$$

To prove this observation, we first define case-based conjectures. For every $i < t \leq T-1$, $x, z \in X$, let

$$A_{i,t,x,z} = \{\omega \in \Omega \mid \omega_X(i) = x, \omega_X(t) = z, \omega_Y(i) = \omega_Y(t)\}$$

and observe that it is the union of finitely many sets of the type $[h_t, Y']$. Hence $A_{i,t,x,z} \in \mathcal{A}$ and $\{A_{i,t,x,z}\} \in \mathcal{E}$.

We can interpret this conjecture as indicating that, *if* the input data are given by x in period i and by z in period t, *then* periods i and t will produce the same outcome (value of y). Notice that in contrast to the Bayesian conjectures, a single case-based conjecture consists of many states: $A_{i,t,x,z}$ does not restrict the values of $\omega_X(k)$ or $\omega_Y(k)$ for $k \neq i, t$. Let the set of all conjectures of this type be denoted by

$$\mathcal{CB} = \{A_{i,t,x,z} \mid i < t \leq T, x, z \in X\} \subset \mathcal{A}. \tag{5}$$

A credence function φ is *case-based* if, for every set $E \in \mathcal{E}$, we have

$$\varphi(E) = \varphi(E \cap \mathcal{CB}). \tag{6}$$

Thus, among the conjectures contained in the set $\mathcal{A}$, only those in $\mathcal{CB}$ are assigned positive weight by a case-based credence function.

Once the set of conjectures $\mathcal{CB}$ has been defined, the proof of Lemma 2 is straightforward:

Proof. Given a similarity function s and $\beta \leq 1$, let

$$\varphi(\{A_{i,t,x,z}\}) = c_t \beta^{(t-i)} s(x,z) \tag{7}$$

where $c_t > 0$ is chosen so that $\varphi(\mathcal{CB})$ is finite, say, $c_t = t^{-2}$. Let, for $E \in \mathcal{E}$,

$$\varphi(E) = \sum_{\{A_{i,t,x,z}\} \in E} \varphi(\{A_{i,t,x,z}\}).$$

Consider a history $h_t = h_t(\omega)$ and a prediction $y \in Y$. To calculate $\varphi(\mathcal{A}(h_t, \{y\}))$ observe first that, at h_t, only the conjectures $\left\{A_{i,t,\omega_X(i),\omega_X(t)} \mid i < t\right\}$ are unrefuted and yield predictions that are included in the singleton $\{y\}$. Hence, only t conjectures will affect the prediction $\{y\}$, corresponding to the t possible case-based conjectures of the form $A_{i,t,\omega_X(i),\omega_X(t)}$ (with $i = 0, 1, ..., t-1$). It is then immediate that $\varphi(\mathcal{A}(h_t, \{y\})) = c_t S(h_t, y)$. ■

In general, we could define similarity relations based not only on single observations but also on sequences, or on other more general patterns of observations, and could define predictions over nonsingleton sets. Such higher-level analogies can also be captured as conjectures in our framework. For instance, the agent might find history h_t similar to history h_i for $i < t$, because in both of them the last k periods had the same observations. This can be reflected by conjectures including states in which observations $(i-k+1), ..., i$ are identical to observations $(t-k+1), ..., t$, and so forth.

3.3 Rule-Based Reasoning

The model can accommodate many other forms of reasoning, often referred to as "rule-based reasoning."[17] These other modes of reasoning are again characterized by conjectures or "rules" to which they attach weights of credence. This section provides some examples.

The rule "the price of oil always rises" corresponds to the conjecture

$$A = \{\omega \in \Omega \mid \omega_Y(t) = 1 \quad \forall t\}.$$

[17]We draw the name "rule-based" from earliest models of reasoning, dating back to Greek philosophy and its study of logic, focusing on the rules of deduction and the concept of proof. The rise of analytical philosophy, the philosophy of mathematics, and artificial intelligence greatly extended the scope of rule-based reasoning, including its use for modeling human thinking, as in the introduction of non-monotonic (McCarthy [24], McDermott and Doyle [25], Reiter [29]), probabilistic (Nilsson [26]), and a variety of other new logics.

There are many states in this conjecture, featuring different sequences of changes in the values of the level of demand and conflict.

Our framework can also encompass *association rules*, or rules that can be expressed as *conditional* statements. For example, consider the rule "if the level of conflict has risen, so will the price of oil." This rule can be described by

$$A = \{\omega \in \Omega \mid \omega_{X^2}(t) = 0 \quad or \quad \omega_Y(t) = 1 \quad \forall t\}. \tag{8}$$

(Recall that $\omega_{X^2}(t)$ indicates whether there was an increase in the index of conflict, and $\omega_Y(t)$ an increase in the price of oil.) The rule "an increase in conflict implies an increase in the price of oil" is then read as "either there will be less conflict, or more expensive oil, or possibly both."[18]

An association rule will be excluded from the set $\mathcal{A}(h_t, Y')$ as soon as a single counter-example is observed. Thus, if history h_t is such that for some $i < t$ we observed an increase in the level of conflict that was not followed by a rise in the price of oil, the conjecture (8) will not be used for further analysis. When an association rule is unrefuted, it may or may not affect predictions, depending on whether its antecedent holds. If the antecedent of a rule is false, the rule becomes vacuously true and does not affect prediction. However, if (in this example) we do observe a rise in the level of conflict, $\omega_{X^2}(t) = 1$, the rule has bite (retaining the assumption that it is as yet unrefuted). Its weight of credence φ will be added to the prediction that the price of oil will rise, $\omega_Y(t) = 1$, but not to the prediction that it will not, $\omega_Y(t) = 0$.

Our framework also allows one to capture functional rules, stating that the value of y is a certain function f of the value of x, such as

$$A = \{\omega \in \Omega \mid \omega_Y(t) = f(\omega_X(t)) \quad \forall t\}.$$

[18]Holland's [18] genetic algorithms address classification problem where the value of y is to be determined by the values of $x = (x^1, ..., x^m)$, based on past observations of x and y. The algorithm maintains a list of association rules, each of which predicts the value of y according to values of some of the x^j's. For instance, one rule might read "if x^2 is 1 then y is 1" and another, "if x^3 is 1 and x^7 is 0 then y is 0." In each period, each rule has a weight that depends on its success in the past, its specificity (the number of x^j variables it involves) and so forth. The algorithm chooses a prediction y that is a maximizer of the total weight of the rules that predict this y and that apply to the case at hand. The prediction part of genetic algorithms is therefore a special case of our framework, where the conjectures are the association rules involved. However, in a genetic algorithm the set of rules does not remain constant, with rules instead being generated by a partly-random process, including crossover between "parent genes," mutations, and so forth.

3.4 Combined Models

The previous subsections illustrate how our framework can capture each of the modes of reasoning separately. Its main strength, however, is in being able to smoothly combine such modes of reasoning, simply by considering credence functions φ that assign positive weights to sets of conjectures of different types.

For example, consider an agent who attempts to reason about the world in a Bayesian way. The agent has a prior probability p over the states of the world, Ω. However, she also carries with her some general rules and analogies. Assume that she employs a model φ such that

$$\varphi(\mathcal{B}) = 1 - \varepsilon$$

(where $\varepsilon > 0$) with weight allocated among the Bayesian conjectures according to

$$\varphi(\{\{\omega\} \mid \omega \in A\}) = (1 - \varepsilon)\, p(A)$$

(for all $A \in \mathcal{A}$) and the remaining weight ε is split among case-based and rule-based conjectures.

If ε is small, then the non-Bayesian conjectures will play a relatively minor role in determining the agent's initial predictions, and will continue to do so as long as the history unfolds along a path that is sufficiently likely under the the prior p. But what happens if the reasoner faces a surprising outcome, such as the September 11 attacks or the Lehman Brothers' collapse?

If the agent had assigned the outcome zero probability, Bayesian updating will not be well-defined. In this case, the non-Bayesian conjectures will determine the agent's predictions. For example, in the face of the September 11 attack, the agent might discard Bayesian reasoning and resort to the general rule that "at the onset of war, the stock market plunges." Alternatively, the agent may resort to analogies, and predict the stock market's behavior based on past cases such as the attack on Pearl Harbor.

Even if the outcome in question had a nonzero but very small prior probability, non-Bayesian reasoning will again be relatively more important. Conditional probabilities are now well-defined and can be used, but the formerly negligible non-Bayesian conjectures will now be much more prominent. This can be interpreted as if the reasoner has a certain degree of doubt about her own probabilistic assessments, captured by the weight $\varepsilon > 0$ put on non-Bayesian conjectures. This process will be formalized in the sequel.

3.5 How Would We Know and Why Would We Care?

We have noted in Section 3.1 that an agent who attaches weight to non-Bayesian conjectures will generate rankings $\succsim_{h_t}$ that are observably non-Bayesian. However, Proposition 1 also notes that the agent's predictions may be consistent with many credence functions φ. Indeed, it is easy to see that, if the agent is asked simply to identify the most likely singleton in Y after each history, then any given sequence of such predictions can be explained by either Bayesian or other methods of reasoning.[19] Why should we care, then, about the mode of reasoning the agent employs?

The answer is that different modes of reasoning might explain a given dataset of predictions *ex post*, yet provide different predictions *ex ante*. For example, if we knew that the agent were Bayesian, we would try to use her past predictions to estimate her prior, and use it to forecast her posterior.[20] By contrast, if the agent were known to be a case-based reasoner, her past predictions would be used to estimate her similarity function. Thus, the same dataset of observations might be compatible with both assumptions about the mode reasoning, but it might lead to different predictions under these assumptions.

This is a manifestation of a more general point: when comparing different paradigms, one often cannot expect to have a simple experiment that identifies the correct one. Within each paradigm many theories may be developed, which can, post hoc, explain given data. However, the simplest theory within one paradigm might lead to rather different predictions than the corresponding theory within another paradigm. In other words, if we augment paradigms with a method for selecting theories within them (say, the simplest theory that fits the data), the choice of a paradigm will have observable implications.

[19] Relatedly, Matsui [23] demonstrated that expected utility maximization and case-based decision theory lead to equivalent sets of feasible outcomes.

[20] Naturally, such a task requires additional assumptions on the structure of the prior probability.

4 Dynamics of Reasoning Methods

4.1 When is Bayesian Reasoning Fragile?

Under what conditions will Bayesian reasoning survive as evidence accumulates, and when will the agent turn to other modes of reasoning? Our answer is that Bayesian reasoning will wither away if the agent's prior is not sufficiently informative.

4.1.1 Assumptions

We start by assuming that at least some weight is placed on both Bayesian and case-based reasoning:

Assumption 1

$$\varphi(\mathcal{B}), \varphi(\mathcal{CB}) > 0.$$

There can be many other types of conjectures that get non-zero weight according to φ. The specific inclusion of case-based reasoning is a matter of convenience, born out of familiarity. We explain in Section 4.1.2 how this assumption could be reformulated to make no reference to case-based conjectures.

Next, we think of the agent as allocating the overall weight of credence in a top-down approach, first allocating weights to modes of reasoning, and then to specific conjectures within each mode of reasoning. First consider the weight of the Bayesian conjectures, $\varphi(\mathcal{B})$. We are interested in an agent who knows relatively little about the process she is observing. An extreme case of such ignorance is modeled by a uniform prior:

$$\frac{\varphi\left(\mathcal{B}(h_t)\right)}{\varphi\left(\mathcal{B}(h'_t)\right)} = 1, \tag{9}$$

for any pair of histories of the same length, h_t and h'_t. We can relax this assumption, requiring only that the probability assigned to any particular history cannot be too much smaller than that assigned to another history of the same length. Thus, one may assume that there exists $M > 1$ such that, for every t and every $h_t, h'_t \in H_t$,

$$\frac{\varphi\left(\mathcal{B}(h_t)\right)}{\varphi\left(\mathcal{B}(h'_t)\right)} < M. \tag{10}$$

We weaken this condition still further, allowing M to depend on t, and assume only that the ratio between the probabilities of two histories cannot go to infinity (or zero) too fast as we consider ever-larger values of t. Formally,

Assumption 2 *There exists a polynomial $P(t)$ such that, for every t and every two histories $h_t, h'_t \in H_t$,*

$$\frac{\varphi(\mathcal{B}(h_t))}{\varphi(\mathcal{B}(h'_t))} \leq P(t).$$

Assumption 2 is still strong—it will be violated if, as is often assumed in Bayesian models, the agent believes she faces successive *iid* draws, say, $\omega_Y(t) = 1$ in each period with probability $p > 0.5$.[21] In this case the agent knows a great deal about the data generating process, being able to identify the process up to the specification of a single parameter. In contrast, our message is that Bayesian reasoning will fade when the agent knows relatively little about the data generating process. However, sub-section 4.1.4 shows that a similar (but somewhat more cumbersome) result holds in the *iid* case as well.

We make an analogous assumption regarding the way that the weight of credence is distributed among the various case-based conjectures. It would suffice for our result to impose a precise analog of Assumption 2, namely that there is a polynomial $Q(t)$ such that, for any t and any pair of case-based conjectures $A_{i,t,x,z}$ and $A_{i',t',x'z'}$, we have

$$\frac{\varphi(\{A_{i,t,x,z}\})}{\varphi(\{A_{i',t',x',z'}\})} \leq Q(t). \tag{11}$$

However, suppose (analogously to (7)) that there exists a similarity function $s : X \times X \to \mathbb{R}_+$, a decay factor $\beta \in (0, 1]$, and a constant $c > 0$ such that, for every $i < t$ and every $x, z \in X$,

$$\varphi(\{A_{i,t,x,z}\}) = c\beta^{t-i}s(x, z). \tag{12}$$

In this case, the characteristics $x, z \in X$ determine the relative weights placed on the case-based conjectures involving information of a given vintage (i.e.,

[21] For an easy illustration of this failure, observe that the ratio of the probabilities of a string of t successive 1's and a string of t successive 0's is $(p/(1-p))^t$, and hence exponential in t.

a given value of $t-i$), with $\beta \leq 1$ ensuring that older information is no more influential than more recent information. This formulation is rather natural, but it violates (11) if $\beta < 1$, as the relevance of older vintages then declines exponentially. Fortunately, there is an obvious and easily interpretable generalization of (11) that allows us to encompass (12).

Assumption 3 *There exists a polynomial $Q(t)$ such that*

- *for every i, i', t, t', x, x' and z, z' with $t - i = t' - i'$, and $t' < t$,*

$$\frac{\varphi(\{A_{i',t',x',z'}\})}{\varphi(\{A_{i,t,x,z}\})} \leq Q(t), \tag{13}$$

- *and for every t, $x, z \in X$ and $i < i' < t$,*

$$\frac{\varphi(\{A_{i,t,x,z}\})}{\varphi(\{A_{i',t,x,z}\})} \leq Q(t). \tag{14}$$

Condition (13) stipulates that within a set of conjectures based on similarities across a given time span (i.e., for which $t - i = t' - i'$), the agent's weights of credence cannot be too different. Condition (14) stipulates that when comparing similarities at a given period t, based on identical characteristics but different vintages, the older information cannot be considered too much *more* important than more recent information. Typically, we would expect older information to be *less* important and hence this constraint will be trivially satisfied.

4.1.2 Result

The following result establishes that under Assumptions 1–3, in the long run the agent puts all of her weight on non-Bayesian (rather than on Bayesian) conjectures.

For the statement of the result we need a notation for the case-based conjectures that are relevant at history h_t:

$$\mathcal{CB}(h_t) = \mathcal{CB} \cap \left(\cup_{y \in Y} \mathcal{A}(h_t, \{y\})\right).$$

Proposition 2 *Let Assumptions 1–3 hold. Then at each $\omega \in \Omega$,*

$$\lim_{t \to \infty} \frac{\varphi\left(\mathcal{B}(h_t)\right)}{\varphi\left(\mathcal{CB}(h_t)\right)} = 0.$$

Hence, the Bayesian component of the agent's reasoning will wither away. As we noted in Section 3.5, the resulting shifting weights of credence can give rise to predictions that could not be rationalized by a Bayesian model.

The Bayesian part of the agent's beliefs converges to the truth at an exponential rate as evidence is accumulated (that is, as t grows): within the Bayesian class of conjectures, the probability of the true state *relative to* the probability of all unrefuted states grows exponentially with t. How is this fast learning reconciled with Proposition 2? The conditional probability of the true state increases at an exponential rate not because its numerator (the weight attached to the true state) increases, but because its denominator (the total probability of all unrefuted states) decreases at an exponential rate. But this is precisely the reason that the weight of the entire class of Bayesian conjectures tapers off and leaves the stage to others, such as the case-based conjectures. As t grows, the weight of Bayesian conjectures that remain unrefuted by history h_t, $\varphi(\mathcal{B}(h_t))$, becomes an exponentially small fraction (given Assumption 1) of the original weight of all Bayesian conjectures, $\varphi(\mathcal{B})$. In contrast, the number of case-based conjectures at period t is only a polynomial (in t), and hence there is no reason for the weight of those that make predictions at history h_t to decrease exponentially fast in t. The *relative* weight placed on Bayesian conjectures thus declines to zero.

It follows that a similar result would hold if we were to replace the class of case-based conjectures with any other class of conjectures that grows polynomially in t and that provides some non-tautological prediction for each h_t, provided an assumption similar to Assumption 3 holds. Therefore, we do not view this result as proving the prevalence of case-based reasoning. Rather, the result highlights the fragility of Bayesian reasoning. Case-based reasoning is simply a familiar example of a mode of reasoning with the requisite properties.

Recall that case-based prediction can be viewed as generalizing the prediction of the modal outcome in the past, as well as the prediction of the most recent outcome. While we again emphasize that the role of case-based reasoning in this argument could be filled by many alternatives, we find it unsurprising that an agent who does not know much about the data generating process may use simple statistical techniques, predicting outcomes that have been observed most often or most recently. Our result describes a possible mechanism by which this may happen, for reasons unrelated to bounded rationality or to cognitive or computational limitations.

4.1.3 Weights of Credence

Proposition 2 is driven by the fact that there are fewer case-based conjectures than there are Bayesian ones. In order for a relatively small class of conjectures to have unrefuted representatives at each history, it must be the case that many of these conjectures make no predictions at many histories. In a sense, conjectures from the smaller class may be viewed as saving their ammunition and picking their fights selectively.

The obvious question is then: Are the Bayesian conjectures treated fairly by our assumptions on the function φ? Specifically, if, at time t, the agent compares the Bayesian conjectures to the case-based ones, she will find that each of the former (that is still in the game) has made t successful predictions, whereas each of the surviving case-based conjectures has made no predictions at all. Shouldn't the tested conjectures get more weight than the untested ones? Shouldn't the credence function φ be updated to reflect the fact that some conjectures have a more impressive track record than others?

Section 2.3 explained that it sacrifices no generality to work with a function φ that is never revised as history unfolds. This simply refocuses the question in terms of the a priori assignment of weights, in the process directing attention to Assumption 3. Should we not make the weight of case-based conjectures of the form $A_{i,t,x,z}$ decline exponentially fast in t (violating Assumption 3), to give the Bayesian ones a fair chance, as it were?

We believe there are some obvious circumstances in which the answer is negative. Suppose that all of the Bayesian conjectures get the same weight, satisfying an extreme version of Assumption 2. It then cannot help but be the case that *some* of them are still unrefuted by history h_t: by construction, there had to be states of the world that are compatible with h_t. The agent knew at time $t = 0$ that, whatever history materializes at time t, some Bayesian conjectures will be in the game. In this case, there is no reason to artificially increase the relative weight of these conjectures upon reaching history h_t, as if they were a priori selected. Adopting the equivalent but a priori convention of decreasing the weight of the case-based conjectures at an exponential or even faster rate strikes us as similarly unjustified, being tantamount to committing to a Bayesian approach that one knows is both tautologically true and without content.[22]

[22] At the other extreme, suppose that only one Bayesian conjecture is given positive weight by φ. In this case, at time t, if this conjecture is still unrefuted, the agent might indeed wish to put an exponentially high relative weight on it, that is, to shrink the total

Another way to look at this problem is the following. Let the agent ask herself at time 0, how much weight is given (a priori) to all conjectures of a given type that will be relevant for prediction at time t. For the Bayesian conjectures the answer is independent of t: when we sum across all possible histories, we always get the same number, $\varphi(\mathcal{B})$, because the union of the relevant conjectures across all histories of length t is the set of all Bayesian conjectures, $\mathcal{B}$, for all t. For the case-based conjectures the situation is quite different: when we consider $t \neq t'$, the set of conjectures that will be relevant at some history h_t is disjoint from the corresponding set for t'. Indeed, we have observed that the total weight of all conjectures that may be relevant at time t has to tend to zero, whereas the corresponding weight for the Bayesian conjectures in a constant. From this viewpoint, the Bayesian conjectures have an inherent advantage. Thus, it seems reasonable to require that, at the very least, the vanishing sequence of weights of case-based conjectures not vanish too fast, and this is what Assumption 3 states.

4.1.4 The *iid* Case

An obvious case in which Assumption 2 is violated occurs when the agent believes that she observes an *iid* process. Suppose, for example, that $Y = \{0, 1\}$ and the agent believes the y_t are *iid* Bernoulli random variables, i.e., $y_t \sim B(p)$. Then Assumption 2 holds only if $p = 0.5$, because the ratio of single states' probabilities involves exponentials of p and $(1-p)$. Nonetheless, a conclusion very similar to that of Proposition 2 still holds.

Consider the state space Ω endowed with the σ-algebra Σ defined by the variables $(x_t, y_t)_{t\geq 0}$. A probability measure μ on Σ is a *non-trivial conditionally iid measure* if, for every $x \in X$ there exists $\lambda_x \in \Delta(Y)$ such that (i) for every $h_t = ((x_0, y_0), \ldots, (x_{t-1}, y_{t-1}), x_t)$, the conditional distribution of Y given h_t according to μ is λ_{x_t}; and (ii) λ_x is non-degenerate for every $x \in X$. The next assumption states that the Bayesian part of the agent's beliefs is

weight of the competing case-based conjectures exponentially fast in t. Equivalently, the agent might arrange at the beginning of the game to cause the weight placed on case-based conjectures $A_{i,t,x,z}$ to decrease very quickly in t, allowing the lone Bayesian conjecture to rule the roost if it survives, while retaining the *relative* weights on the surviving case-based conjectures so that their predictions are unaffected in the event the Bayesian hypothesis is falsified. Notice, however, that this manipulation is unnecessary. If the initial weight attached to Bayesian hypotheses is large, the weight will remain large, as there are no falsified Bayesian hypotheses to melt away. In this case, Bayesian reasoning survives even without help in the form of declining case-based weights.

governed by such a measure:

Assumption 4 *There exists a non-trivial conditionally iid measure μ such that, for every $A \in \Sigma$*

$$\varphi\left(\{\{\omega\} \mid \omega \in A\}\right) = \mu(A)\varphi(\mathcal{B})$$

Thus, this assumption states that the weight of the Bayesian conjectures, $\varphi(\mathcal{B})$, is divided among them in a way that is proportional to the measure μ.[23]

We can now state

Proposition 3 *Let Assumptions 1,3, and 4 hold, and let μ be the measure of Assumption 4. Then*

$$\mu\left(\lim_{t\to\infty} \frac{\varphi\left(\mathcal{B}(h_t)\right)}{\varphi\left(\mathcal{CB}(h_t)\right)} = 0\right) = 1.$$

Proposition 3 states that, μ-almost surely, the weight of the Bayesian hypotheses relative to that of the case-based ones will converge to zero. Thus, an agent who has Bayesian beliefs μ, and who puts some weight $\varepsilon > 0$ on the case-based beliefs in a way that corresponds to Assumption 3, will, *according to her own beliefs*, converge to be non-Bayesian. Importantly, even if the agent were right about the Bayesian part of her beliefs, she would still predict that her own reasoning will become non-Bayesian.

The proof of Proposition 3 mimics that of Proposition 2. The key observation is that there are exponentially many histories for any given frequencies of outcomes, provided that these frequencies are non-trivial. For example, if $|X| = 1$, $Y = \{0, 1\}$ and we consider a history of length t, there is but one history in which there are 0 y's that are equal to 1, and $O(t^k)$ histories in which there are k such y's. But there are exponentially many histories in which the relative frequency of 1 is close to pt. That is, $\binom{t}{pt}$ is exponential in t if $p \neq 0, 1$. More generally, since μ is assumed to be *non-trivial* conditionally *iid* (that is, since the conditional distributions λ_x are assumed to be non-degenerate), apart from a set of μ measure zero, any history at time t has exponentially many other histories that are just as likely.

[23] Observe that μ is defined over subsets of Ω (that are in Σ) whereas φ is defined over subsets of such subsets, and the assumption only deals with the subsets that contain only singletons $\{\omega\}$. Observe also that Assumption 4 remains silent about the distribution of the x's.

Observe that a similar result would hold in case the agent only believes that the variables y_t, given each value of x, are exchangeable. Indeed, the very definition of exchangeability, involving all possible permutations, hints at the dangers of exponential blow-up. To prove such a result, one need only make sure that sufficiently many permutations result in different histories h_t.

Along similar lines, the conclusion of Proposition 2 holds also if Assumption 2 fails but there exists $\gamma < 1$ such that, for some polynomial $P(t)$, for every t and every h_t,

$$\varphi(\mathcal{B}_t(h_t)) \leq \gamma^t P(t). \tag{15}$$

If Assumptions 1 and 3 also hold, then the relative weight on Bayesian conjectures will decline to zero.

On the other hand, an agent who is *resolutely* Bayesian will handle *iid* variables just as one would expect.

Example 1 Suppose that X is degenerate, say, $X = \{0\}$, $Y = \{0, 1\}$, and that $\varphi(\mathcal{B}) = 1$. Suppose this Bayesian's prior is a nondegenerate mix of two prior distributions. One of these predicts that each y_t is drawn from the Bernoulli distribution $B(p)$ and one from the Bernoulli distribution $B(q)$. Then if the observations are indeed *iid* draws from $B(p)$, the relatively weight that the agent places on this distribution will almost surely (according to the true distribution) converge to one. The agent will eventually predict optimally, predicting $y_t = 1$ in each period if $p > .5$, and $y_t = 0$ otherwise. ■

Agents who are convinced they should be Bayesian thus exhibit familiar behavior. In light of this, how will an agent reason who is reasonably confident that she faces an *iid* process, but is uncertain about the parameter? Much depends on what is meant by "reasonably confident," a judgment manifested in the agent's credence function.

Example 2 Suppose that X is degenerate, $Y = \{0, 1\}$, and that the weights of credence the agent attaches to Bayesian conjectures are consistent with the various y_t being drawn *iid* from a Bernoulli distribution. If the agent is sufficiently confident of her *iid* hypothesis as to attach zero weights of credence to all other conjectures, she will be familiarly and resolutely Bayesian throughout. If she hedges her bets by attaching some weight to case-based conjectures, in accordance with Assumption 3, then the weight she attaches to Bayesian conjectures will decline to zero. Indeed, the agent can be assured

at the beginning of the process that this will happen. Alternatively, suppose the agent attaches credence to case-based hypotheses, but attaches weights that cause $\varphi(\{A_{i,t,x,z}\})$ to decline exponentially (violating Assumption 3). Suppose again the Bayesian prior is a nondegenerate mix of two prior distributions, one predicting that each y_t is drawn from the Bernoulli distribution $B(p)$ and one predicting draws from the Bernoulli distribution $B(q)$. Then if the weights $\varphi(\{A_{i,t,x,z}\})$ decline appropriately, the weight of Bayesian hypotheses will increase to one if the data are indeed generated by $B(p)$ or $B(q)$, but in other cases (e.g., data drawn from $B(r)$ for some $r \neq p, q$) the Bayesian prior will slip into insignificance and case-based reasoning will prevail. They agent will thus remain Bayesian if her Bayesian prior contains the correct date generating process, but will otherwise slip into case-based reasoning. ■

Example 2.5 We can provide a simple illustration. Consider again the simplest case of $X = \{0\}$, $Y = \{0, 1\}$. Assume that y_t are *iid*, where $y_t = 1$ with probability p.

Consider the set of states

$$B_{i,y} = \{\omega \in \Omega \mid \omega_Y(t) = y \quad \forall t \geq i\}$$

for $i \geq 0$ and $y \in Y$. Hence, each collection of states $B_{i,y}$ is identified by a given period i and outcome y, and predicts that from period i on, only outcome y will be observed.

The agent attaches weights to all of the Bayesian conjectures in the set

$$\bigcup_{i \geq 0, y \in Y} B_{i,y}.$$

The weight attached to the states in $B_{i,0} \cup B_{i,1}$ is given by $\xi 2^{-i}$, evenly distributed among such states. (The factor ξ is a normalization, to ensure weights sum to unity.) There are 2^i such hypotheses, so that the weight of a single such hypothesis is 2^{-2i}. Notice that (15) holds, so that the weights attached to Bayesian hypotheses are consistent with Bayesian reasoning withering away.

If weight is attached only to Bayesian hypotheses, then obviously Bayesian reasoning will survive. If Assumption 3 holds, then Bayesian reasoning will disappear. We examine a middle case here.

Because there are no x values to consider, the case-based conjectures are simply

$$A_{i,t} = \{\omega \in \Omega \mid \omega_Y(i) = \omega_Y(t)\},$$

and the set of all case-based conjectures is

$$\mathcal{CB} = \{A_{i,t} \mid i < t\}.$$

The weight attached to the case-based hypothesis in $A_{i,t}$ is given by $\xi 2^{-2t}$ (obviating Assumption 3).

Now consider a history h_t ending with a run of either $y = 0$ of $y = 1$ of length ℓ. The total weight of case-based hypotheses is 2^{-2t}. Among the Bayesian hypotheses, there survives a hypothesis from the $B(t-\ell, 0) \cup B(t-\ell, 1)$, with weight $2^{-(t-\ell)}$. The weight placed on Bayesian hypotheses is thus at least $2^{-(t-\ell)}$, and the weight of Bayesian hypotheses relative to case-based hypotheses is at east 2^{ℓ}. As the history unfolds, with probability one there will occur arbitrarily long strings of identical values of y, at which point the relative weight of Bayesian hypotheses will be arbitrarily large. At the same time, strings will periodically be broken, restoring the Bayesian and case-based hypotheses to an equal footing. Hence, from time to time there will emerge a Bayesian hypotheses that is accepted, only to collapse at some subsequent point. In other words, even if the data are completely random, it should be expected that theories would rise and fall every so often, with case-based reasoning being more prominent between regimes of different theories.

Observe that the balance of weights between the two modes of reasoning is driven by the success of Bayesian reasoning. This reflects the intuition that people would like to understand the process they observe, and that such "understanding" means a simple, concise theory that explains the data. If such a theory exists, agents will tend to prefer it over case-based reasoning. But when all simple theories are refuted, agents will resort to case-based reasoning. ■

4.2 When will Bayesianism Prevail?

If we had worked with the stronger version of Assumption 2 given by (9), we would have the expected (though perhaps reassuring) result that Bayesian reasoning disappears when the Bayesian prior is so diffuse that a Bayesian could not possibly learn anything. However, Assumption 2 allows Bayesian priors that will in turn allow learning, and yet still give way to other sorts

of reasoning. Section 4.1.4 shows that Assumption 2 can be weakened yet further. When will Bayesian reasoning remain useful in the long run, or even dominate other reasoning methods?

Example 3 Suppose the agent believes that she nearly knows the true state of the world. We capture this by letting there be some ω, $\varphi(\{\omega\}) = 1 - \varepsilon$ (and hence allowing Assumption 2 to fail). If, on top of this, the agent is also correct in her focus on state ω, then (that is, at state ω) the weight attached to Bayesian conjectures will never dip below $1 - \varepsilon$. In other words, if the agent believes she knows the truth, and *happens to be right*, her Bayesian beliefs will remain dominant. ■

Example 4 A slightly less trivial example is the following. Suppose the agent believes she faces a cyclical process, but is uncertain of its period. To capture these beliefs in a simple model, let us consider only Bayesian and case-based reasoning. In addition, let $X = \{0\}$ and $Y = \{0, 1\}$, so that all periods have the same observable features, and they only differ in the binary variable the agent is trying to predict. For $k \geq 1$, let $\omega^k \in \Omega$ be defined by

$$\omega^k_Y(t) = \begin{cases} 0 & 2mk \leq t < (2m+1)k \quad m = 0, 1, 2, ... \\ 1 & (2m+1)k \leq t < (2m+2)k \quad m = 0, 1, 2, ... \end{cases}.$$

Thus, for $k = 1$ the process is 01010101..., for $k = 2$ it is 001100110011... and so forth.

Let the agent's beliefs satisfy

$$\varphi(\{\{\omega^k\}\}) = \frac{1 - \varepsilon}{2^k}$$

and

$$\varphi_T(\{\{\omega\} | \omega \notin \{\omega^k \,|1 \leq k\}\}) = 0.$$

Thus, the agent splits all the weight of the Bayesian conjectures among the conjectures $\{\omega^k\}$ and leaves no weight to the other Bayesian beliefs.[24] Once again, Assumption 2 fails. The remaining weight, ε, is split among the case-based conjectures.

[24] Observe that these Bayesian beliefs can also be readily described as rule-based beliefs. We suspect that this is not a coincidence. When Bayesian beliefs violate Assumption 2, it is likely to be the case that they reflect some knowledge about the data generating process, which can also be viewed as believing in a class of rules.

Next suppose that the agent is right in her belief that the process is indeed cyclical (starting with a sequence of 0's). Thus, the data generating process chooses one of the states ω^k. At this state, once we get to period $t = k$, all the Bayesian conjectures $\left\{\omega^{k'}\right\}$ for $k' \neq k$ are refuted. In contrast, the conjecture $\left\{\omega^k\right\}$ is not refuted at any t. Consequently, at ω^k, for every $t \geq k$, the total weight of the Bayesian conjectures remains $\frac{1-\varepsilon}{2^k}$. The total weight of the case-based conjectures converges to 0, resulting in the Bayesian mode of reasoning remaining the dominant one (for large t). Clearly, this will only be true at the states $\left\{\omega^k\right\}$. At other states the converse result holds, because all Bayesian conjectures will be refuted and case-based reasoning will be the only remaining mode of reasoning. ■

Example 5 Considering the same set-up, $X = \{0\}$ and $Y = \{0, 1\}$, let us limit attention to the first T periods. Consider a Bayesian agent who has a uniform belief over the average

$$\bar{y}_T = \frac{1}{T}\sum_{t=0}^{T-1} \omega_Y(t)$$

and, given $\bar{y}_T$, a uniform distribution over all the corresponding states. Thus, the agent puts a weight of $\frac{1}{T+1}$ on the sequence $1, 1, ..., 1$, but only a weight of $\frac{1}{T(T+1)}$ on each sequences with $(T-1)$ 1's and a single 0, and a weight $o(T^{-3})$ on each sequence with two 0's, and so forth.

The total weight of all case-based conjectures is a convergent series. This implies that the weight of all the case-based conjectures that are relevant at T has to decline to zero at a rate that is faster than $\frac{1}{T}$. Hence, if the agent observes the sequence $1, 1, ..., 1$, she will put more weight on the Bayesian conjecture that can be described also by the rule "$\omega_Y(t) = 1$ for every t." However, if the agent observes one exception to this rule, the Bayesian conjecture that predicts only 1's will have a weight that is $o(T^{-2})$. The more exceptions one observes, the lower is the weight of the Bayesian conjectures.

If the rate of decline of the weight of case-based conjectures in polynomial in T, say, $o(T^{-k})$ for $k > 1$, then finitely many exceptions to the rule "y is always 1" will suffice to switch to case-based reasoning. (Observe, however, that this reasoning is likely to make similar predictions: if all but k times one has observed $y_t = 1$, the modal prediction will still be $y_T = 1$.) If, by contrast, the weight of case-based conjectures decreases exponentially fast in

T, even very spotty patterns will keep the Bayesian conjectures on par with the case-based ones. ■

In summary, for Bayesian reasoning to prevail, the reasoner's Bayesian beliefs must be sufficiently informative (i.e., must contain the truth and must not be too diffuse), and the reasoner must have sufficient confidence in those Bayesian beliefs (e.g., build quickly declining weights of credence into the case-based conjectures relevant to successive periods). Economic models typically ensure this confidence by assuming that the agent entertains only Bayesian conjectures. We emphasize that our purpose is *not* to criticize either Bayesian reasoning or models based on Bayesian reasoning. Rather, our point is that the same characteristics that make Bayesian reasoning work well for a committed Bayesian can make it fragile in the hands of a tentative Bayesian.

4.3 Probabilistic Reasoning

Our main result establishes conditions under which Bayesian reasoning is fragile. However, it does not imply that when the weight of the Bayesian conjectures becomes negligible (relative to the weight of all unrefuted conjectures), probabilistic reasoning will be forever discarded. Instead, case-based and rule-based reasoning may subsequently give way to a new probabilistic model. Specifically, if at history h_t $\varphi(\mathcal{B}(h_t))$ is low (relative to, say, $\varphi(\mathcal{CB}(h_t))$), it is still possible that at a certain continuation of h_t, $h_{t'}$ with $t' > t$, the agent will again form beliefs that put a high weight on singleton hypotheses consistent with $h_{t'}$. As Proposition 1 indicates, this process is consistent with a single, history-independent φ.

The dynamics of our model can thus capture the type of reasoning raised in Section 1. Given a certain history $h_{t'}$ above, the agent forms probabilistic beliefs that can be thought of as a Bayesian model *given* $h_{t'}$. Such beliefs cannot be guaranteed to assign a high probability to all eventualities. The agent may have failed to seriously consider certain black swans, and some of them will have very low probability. As a result, the agent may find herself at a point where she mistrusts her model, and resorts to case-based and rule-based reasoning in its stead. However, at some subsequent $t'' > t'$, the agent beliefs may again effectively form a new probabilistic model.

It is easier to generate such "conditionally Bayesian" models than a single, a priori Bayesian model, for two related reasons: first, a model that starts at

a history h_t has to consider only a subset of the events that a comprehensive Bayesian model deals with. If the time horizon is finite, the number of states one needs to assign probability to decreases exponentially fast as t grows, which means that the probability assignment task becomes easier. Second, as t grows, there are more data on the basis of which such probability assignments may be done. Indeed, if one considers the Bayesian model, one has to assign probabilities to many states with no data at all, out of thin air as it were. By contrast, for sufficiently large t, the agent may find regularities in the data that may suggest a new probabilistic model for the remaining periods.

At the same time, any probabilistic model generated after some history h_t will eventually face the same difficulty: whatever the finite history used for its formulation, it will become negligible relative to the size of the state space as one looks further into the future. Hence, one should expect that, apart from simple statistical problems, no probabilistic model will ever be the "correct" one. Rather, the agent will be cycling between periods in which she has a satisfactory probabilistic models, and periods in which black swans are observed and model uncertainty reigns. In such periods, case-based and rule-based reasoning are needed to make predictions, and, eventually, to formulate new probabilistic models.

To illustrate this, consider again the simplest example with no predicting variables, say, $X = \{0\}$, and $Y = \{0, 1\}$. Suppose that the agent believes that the data generating process on $\{0, 1\}^{\mathbb{N}}$ follows a probabilistic model given by a measure ρ (defined on the standard σ-algebra on $\{0, 1\}^{\mathbb{N}}$). However, certain periods in the past might have been exceptional – say, periods of wars, financial crises, and so forth. Hence, the agent does not believe that ρ is necessarily the appropriate probability measure to be assigned to Ω. She only believes that *after a certain history*, the continuation of the process will be governed by ρ. In other words, ρ is the conditional belief on Ω given a history h_t (with conditional state space that is also $\{0, 1\}^{\mathbb{N}}$).

Assume further that the agent does not presume that she can assign probabilities to the initial period, in which wars, financial crises, and the like disrupt her prediction. She does not pretend to have probabilistic beliefs over the length of time at which the process will finally stabilize and be govern by φ. Rather, she awaits to see periods of relative calm, in which $y = 1$, and she assigns weight to rules of the type "if $y_i = 1$ for the last k periods, we may finally see the periods governed by ρ."

Let $R_{k,s}$ stand for the conjecture that, after the first time in which k

consecutive 1's were observed, the process will follow a state $s \in \{0,1\}^{\mathbb{N}}$. Explicitly,

$$R_{k,s} = \left\{ \omega \in \Omega \;\middle|\; \begin{array}{c} \exists t \\ \omega_Y(i) = 1 \quad t-k \leq i < t \\ \omega_Y(i) = s(i-t) \quad t \leq i \\ (\omega_Y(i))_{i<t-1} \text{ does not contain a sequence of } k \text{ 1's} \end{array} \right\}.$$

(Observe that $R_{k,s} \cap R_{k',s'}$ may be non-empty for $s \neq s'$ if $k \neq k'$. However, $s \neq s'$ does imply that $R_{k,s}$ and $R_{k,s'}$ are disjoint for every k.)

Let us assume that the agent assigns to the conjectures $\{R_{k,s}\}_s$ a total weight $Q(k) = k^{-2}$ and that ρ, when applied to Ω, satisfies 2. The weight of the "rule-based" conjectures will be given by

$$\varphi(\{R_{k,s}\}_{s \in A}) = Q(k)\rho(A)$$

for a measurable $A \subset \{0,1\}^{\mathbb{N}}$.

Apart from these conjectures, the agent will be assumed to assign positive weight also to the case-based conjectures,

$$A_{i,t} = \{\omega \in \Omega \mid \omega_Y(i) = \omega_Y(t)\},$$

and assume that their weight is

$$\varphi(\{A_{i,t}\}) = Q(t)/t$$

that is, that the total weight of the case-based conjectures at time t is $Q(t) = t^{-2}$ and that this weight is divided equally among the periods preceding t. In particular, Assumption 3 holds.

Consider a history h_t. Every sequence of 1's in it initiates a rule $R_{k,s}$. However, since Assumptions 2 and 3 hold, the probabilistic reasoning embodied by the rules $\{R_{k,s}\}_{s,k \leq t}$ will give way to the case-based reasoning as in Proposition 2. To be precise, fix t and consider the reasoning at history $h_{t'}$ with $t' > t$. There are up to t different sets of conjectures $\{R_{k,s}\}_{s \in \{0,1\}^{\mathbb{N}}}$ that affect the agent's reasoning in a probabilistic way, but since the weight of each of them decreases exponentially fast with t', in the long run the weight of all of them combined will be negligible relative to the weight of the case-based conjectures.

However, as t' grows, new sets of conjectures $\{R_{k,s}\}_{s \in \{0,1\}^{\mathbb{N}}}$ might join the game. Assume that t is large and that the longest sequence of 1's in

h_t is of length $k << t$. When, at some point, $k+1$ 1's are observed, the weight of the probabilistic reasoning in $\{R_{k+1,s}\}_{s\in\{0,1\}^{\mathbb{N}}}$ will be $Q(k+1)$, and it will overwhelm the weight of the case-based reasoning, $Q(t)$. Moreover, this phenomenon is likely to recur. Suppose, for the sake of the argument, that the true data generating process of y_t is uniform (and iid). Considering a given length k, we may ask when will a sequence of k 1's will first appear. With very high probability, this will occur at time t which is exponentially larger than k. At such a history h_t, the agent will reason mostly by the probabilistic reasoning ρ. Over time, when $t' > t$ grows, this probabilistic reasoning will decline in weight, but, with probability 1, new probabilistic models will be developed later on.

Thus, our framework may capture non-trivial dynamics between case-based and rule-based reasoning. Moreover, it can describe how probabilistic theories may be re-developed as history unfolds. Under the assumptions of Proposition 2, a prior distribution that has been formed at time $t = 0$ will have a negligible effect on reasoning in the long term. Yet, conditional probabilistic models might be re-formulated, capturing the agents beliefs that she can make probabilistic predictions from a certain time on.

5 Concluding Remarks

5.1 Methods for Generating Conjectures

In many examples ranging from scientific to everyday reasoning, it may be more realistic to put weight φ not on specific conjectures A, but on methods or algorithms that generate them. For example, linear regression is one such method. When deciding how much faith to put in the prediction generated by the OLS method, it seems more plausible that agents put weight on "whatever the OLS method prediction came out to be" rather than on a specific equation such as "$y_t = 0.3 + 5.47x_t$."

One simple way to capture such reasoning is to allow the carriers of weight of credence to be sets of conjectures, with the understanding that within each set a most successful conjecture is selected for prediction, and that the degree of success of the set is judged by the accuracy of this most successful conjecture. The following example illustrates.

Suppose that the agent is faced with a sequence of datasets. In each dataset there are many consecutive observations, indicating whether a comet

has appeared (1) or not (0). Different datasets refer to potentially different comets.

Now assume that the agent considers the general notion that comets appear in a cyclical fashion. That is, each dataset would look like

$$0, 0, ..., 0, 1, 0, 0, ..., 0, 1, ...$$

where a single 1 appears after k 0's precisely. However, k may vary from one dataset to the next. In this case, the general notion or "paradigm" that comets have a cyclical behavior can be modeled by a set of conjectures—all conjectures that predict cycles, parametrized by k. If many comets have been observed to appear according to a cycle, the general method, suggesting "find the best cyclical theory that explains the observations" will gain much support, and will likely be used in the future. Observe that the method may gain credence even though the particular conjectures it generates differ from one dataset to the next.

5.2 Single-Conjecture Predictions

This paper is concerned with reasoning that takes many conjectures into account and aggregates their predictions. Alternatively, we may consider reasoning modes that focus on a most preferred conjecture (among the unrefuted ones) and make predictions based on it alone. For example, if we select the simplest theory that is consistent with the data, we obtain Wittgenstein's [40] definition of induction.[25] If, by contrast, we apply this method to case-based conjectures, we end up with nearest-neighbor approaches (see Cover and Hart [6] and Fix and Hodges [10, 11]) rather than with the case-based aggregation discussed here.

[25]See Solomonoff [38], who suggested to couple this preference for simplicty with Kolmogorov complexity measure to yield a theory of philosophy of science. Gilboa and Samuelson [12] discuss the optimal selection of the preference relation over theories in this context.

6 Appendix A: Proofs

6.1 Proof of Proposition 1

Define, for each h_t and for every $Y' \subsetneq Y$,

$$\varphi\left(\{[h_t, Y'] \cup [h_t]^c\}\right) = c_{h_t}\varphi_{h_t}(\mathcal{A}(h_t, Y'))$$

for every conjecture of the form $\{[h_t, Y'] \cup [h_t]^c\}$, and set $\varphi(\mathcal{F}) = 0$ where $\mathcal{F}$ is the set of all conjectures that are not of this form, and $c_{h_t} > 0$ is to be determined. Observe that the conjecture $[h_t, Y'] \cup [h_t]^c$is unrefuted (because it contains $[h_t, Y']$, and hence is consistent with the observed history)and non-tautological (because $Y' \subsetneq Y$)only at h_t. Hence, at history h_t, only conjectures of the form $[h_t, Y''] \cup [h_t]^c$ (with $Y'' \subsetneq Y$) are unrefuted and non-tautological, and the total weight that they assign to a subset of outcomes Y' is by construction $c_{h_t}\varphi_{h_t}(\mathcal{A}(h_t, Y'))$. The coefficient c_{h_t} is chosen so that the total weight assigned by φ to all conjectures converges, which would be the case, for instance, if

$$\sum_{h_t \in H_t} c_{h_t} = t^{-2}.$$

■

6.2 Proof of Proposition 2

We start by showing that, because the ratio of weights assigned to specific histories of the same length t is bounded by a polynomial of t, the weight of each particular such history is bounded by this polynomial divided by an exponential function of t.

Consider a period t and a history h_t. If $\varphi(\mathcal{B}(h_t)) > \eta$, then, since for every $h_t, h'_t \in H_t$, $\varphi(\mathcal{B}(h_t)) \leq P(t)\varphi(\mathcal{B}(h'_t))$, for every h'_t,

$$\varphi(\mathcal{B}(h'_t)) \geq \frac{\varphi(\mathcal{B}(h_t))}{P(t)} > \frac{\eta}{P(t)}$$

Observe that $|H_t| \geq d^t$ for $d = |X||Y| > 1$. Hence

$$\varphi(\mathcal{B}) > \frac{d^t\eta}{P(t)}$$

and $\varphi(\mathcal{B}) < 1$ implies

$$\eta < \frac{P(t)}{d^t}$$

Since this is true for every η such that $\eta < \varphi(\mathcal{B}(h_t))$, we conclude that

$$\varphi(\mathcal{B}(h_t)) \leq \frac{P(t)}{d^t}. \tag{16}$$

We now turn to discuss the weight of the case-based conjectures that are relevant for prediction at h_t. We wish to show that this weight cannot be too small. First, observe that the set of case-based conjectures is countable. Denote the total weight of the case-based conjectures whose second period is τ by S_τ. Explicitly,

$$S_\tau = \sum_{i=0}^{\tau-1} \sum_{x,z \in X} \varphi\left(\{A_{i,\tau,x',z'}\}\right)$$

Then,

$$\varphi\left(\mathcal{CB}\right) = \sum_{\tau=1}^{\infty} S_\tau.$$

Choose T large enough so that

$$\sum_{\tau=1}^{T} S_\tau > \frac{\varphi\left(\mathcal{CB}\right)}{2}. \tag{17a}$$

¿From now on, assume that $t \geq T$.

Consider a conjecture $A_{(t-1),t,x,z} \in \mathcal{CB}$ and assume that $\varphi\left(\left\{A_{(t-1),t,x,z}\right\}\right) < \xi$. By (13) (of Assumption 3) we have that, for all $t' < t, x', z'$

$$\varphi\left(\left\{A_{(t'-1),t',x',z'}\right\}\right) < \xi Q(t).$$

By (14) (of that Assumption), we know that for all $i < t' < t$, and all x', z',

$$\varphi\left(\{A_{i,t',x',z'}\}\right) < \varphi\left(\left\{A_{(t'-1),t',x',z'}\right\}\right) Q(t) < \xi \left[Q(t)\right]^2.$$

The overall number of case-based conjectures whose second period is $t' \leq t$ is $|X|^2 \begin{pmatrix} t \\ 2 \end{pmatrix}$. Since the weight of each is less than $\xi \left[Q(t)\right]^2$ we conclude that

their total weight satisfies

$$\sum_{\tau=1}^{T} S_\tau < \xi \left[Q(t)\right]^2 |X|^2 \left(\begin{array}{c} t \\ 2 \end{array} \right)$$

and, using (17a) we obtain

$$\frac{\varphi\left(\mathcal{CB}\right)}{2} < \sum_{\tau=1}^{T} S_\tau < \xi \left[Q(t)\right]^2 |X|^2 \left(\begin{array}{c} t \\ 2 \end{array} \right).$$

Define

$$R(t) = 2 \left[Q(t)\right]^2 |X|^2 \left(\begin{array}{c} t \\ 2 \end{array} \right)$$

and observe that it is a polynomial in t.

Thus, we have

$$\xi > \frac{\varphi\left(\mathcal{CB}\right)}{R(T)}.$$

Since this holds for any ξ such that $\xi > \varphi\left(\left\{A_{(t-1),t,x,z}\right\}\right)$, it has to be the case that

$$\varphi\left(\left\{A_{(t-1),t,x,z}\right\}\right) \geq \frac{\varphi\left(\mathcal{CB}\right)}{R(t)}.$$

We observe that at h_t there are precisely t case-based conjectures that are unrefuted and non-tautological, and among them there is one of the type $A_{(t-1)t,x,z}$ (that is, the one defined by $x = \omega_X(t-1)$ and $z = \omega_X(t)$). It follows that

$$\varphi\left(\mathcal{CB}(h_t)\right) \geq \varphi\left(\left\{A_{(t-1),t,x,z}\right\}\right) \geq \frac{\varphi\left(\mathcal{CB}\right)}{R(t)}. \tag{18}$$

Combining (16) and (18) we obtain

$$\frac{\varphi\left(\mathcal{B}(h_t)\right)}{\varphi\left(\mathcal{CB}(h_t)\right)} < \frac{P(t)R(t)}{\varphi(\mathcal{CB})d^t}$$

where the expression on the right clearly converges to 0 as $t \to \infty$.

■

7 Appendix B: Belief Functions

This appendix is devoted to a brief introduction to Dempster-Shafer's Theory of Belief Functions, as well as its relationship to Choquet expected utility. It contains some definitions and well-known facts, as well as one result (Proposition 4) which is, to the best of our knowledge, new.

7.1 Capacities and Qualitative Capacities

Let there be a finite and non-empty set Y. Recall that in our model, Y denotes the outcomes that may be observed at each period.

The space of set functions on Y is defined by

$$V = \left\{ v : 2^Y \to \mathbb{R} \mid v(\varnothing) = 0 \right\}.$$

A *capacity* is a set function $v \in V$ that is *monotone* with respect to set inclusion, that is, that satisfies $v(A) \leq v(B)$ whenever $A \subset B$ (for $A, B \subset Y$), and that satisfies $v(Y) = 1$.

A binary relation $\succsim \subset 2^Y \times 2^Y$ is a *qualitative capacity* if (where $\succ$ denotes the asymmetric part of $\succsim$)

(i) $\succsim$ is a weak order;

(ii) $\succsim$ is monotone with respect to set inclusion, that is, $B \succsim A$ whenever $A \subset B$;

(iii) $\succsim$ is a non-trivial: $Y \succ \varnothing$.

A relation $\succsim$ is *persistently monotone*, or *p-monotone* for short, if, for every $S \subset T \subset Y$, and every $R \subset Y$ such that $R \cap T = \varnothing$, if $S \cup R \succ S$, then $T \cup R \succ T$. This condition states that, if an event R makes a non-null marginal contribution to another event S, R will make a non-null marginal contribution to any larger event T (in the sense of set inclusion). Thus, p-monotonicity has the flavor of a condition of non-decreasing marginal contribution, stated in the ordinal language, which only allows us to distinguish between null and non-null marginal contributions. Observe that if one has a quantitative measure of marginal contributions by a capacity v, the non-decreasing marginal contribution condition would say that, for every S, T, R as above, $v(T \cup R) - v(T) \geq v(S \cup R) - v(S)$, which is the familiar condition called *convexity*, *super-modularity*, or *2-monotonicity*.

Another way to view p-monotonicity is provided by the following lemma.

Lemma 3 *Let $\succsim$ be a qualitative capacity. The following are equivalent*

(i) $\succsim$ is p-monotone: For every $S \subset T \subset Y$, and every $R \subset Y$ such that $R \cap T = \varnothing$, if $S \cup R \succ S$, then $T \cup R \succ T$.

(ii) For every $A \subset B \subset Y$, for every $C \subset Y$, if $B \cap C \succ A \cap C$, then $B \succ A$.

(iii) For every $D \subset A \subset B \subset Y$, if $D \cup (B \backslash A) \succ D$, then $B \succ A$.

(iv) For every $A \subset B \subset \Omega$, for every $E \subset B$, if $E \succ A \cap E$, then $B \succ A$.

Proof. (i) implies (ii): Assume that $\succsim$ is p-monotone and let there be given $A \subset B$ and C. Define $S = A \cap C$, $T = A$, and $R = (B \cap C) \backslash A$. Then $S \subset T$, $R \cap T = \varnothing$. Since $S \cup R = B \cap C \succ A \cap C = S$ we should also have $T \cup R \succ T$ – where $T \cup R = A \cup (B \cap C)$ and $T = A$. By monotonicity, $B \succsim A \cup (B \cap C)$ and thus $B \succ A$.

(ii) implies (iii): Given $D \subset A \subset B \subset Y$, set $C = D \cup (B \backslash A)$, so that $B \cap C = D \cup (B \backslash A)$ and $A \cap C = D$. Since $D \cup (B \backslash A) \succ D$, we have $B \cap C \succ A \cap C$, and $B \succ A$ follows.

(iii) implies (iv): Given $A \subset B \subset \Omega$ and $E \subset B$, define $D = A \cap E$. We know that $E \succ D$. Clearly, $E \subset D \cup (B \backslash A)$. Hence, by monotonicity, $D \cup (B \backslash A) \succsim E \succ D$ and $B \succ A$ follows.

(iv) implies (i): Given $S \subset T \subset Y$ and $R \subset Y$ with $R \cap T = \varnothing$, define $A = T$, $E = S \cup R$, and $B = T \cup R$. Then $A \cap E = S$. We know that $S \cup R \succ S$ and thus $E \succ A \cap E$. (iv) implies that $B \succ A$, that is, $T \cup R \succ T$. ■

In all three conditions, (ii)-(iv), we have $A \subset B$ and thus, by monotonicity, we know that $B \succsim A$. The additional condition of p-monotonicity says that if there is some evidence that $B \backslash A$ might occur—as implied by the antecedent of (ii), (iii), or (iv)—then B has to be strictly more likely than A.

7.2 Möbius Transforms and Belief Functions

Given a set function $v \in V$, we can define another set function by

$$\phi_v(A) = \sum_{B \subset A} (-1)^{|A|-|B|} v(B).$$

The set function ϕ_v is referred to as the *Möbius transform of* v. We can represent a set function $v \in V$ by its Möbius transform as as follows: for

every $A \subset Y$,

$$v(A) = \sum_{B \subset A} \phi_v(B).$$

Clearly, the mapping $M : V \to V$ defined by $M(v) = \phi_v$ is linear bijection from V onto itself, and so is its inverse M^{-1}. In particular, if a set of Möbius transforms C is convex, so is its corresponding set of set functions $M^{-1}(C)$ and vice versa.

A capacity v is a *belief function* if and only if its Möbius transform is non-negative. Alternatively, one may define a set function v to be *totally monotone* if

$$v(A) \geq \sum_{B \subsetneq A} (-1)^{|A|-|B|} v(B)$$

for all $A \subset Y$, and define a belief function to be a totally monotone capacity.

The bijection of the Möbius transform was first studied, to the best of our knowledge, by Shapley [36].[26] Dempster [9] introduced the concept of a *belief function* to capture the degree of belief in events without requiring additive probabilities (see also Shafer [35]). In Dempster's interpretation, $v(A)$ is a measure of the total belief in an event A, whereas $\phi_v(A)$ is a measure of the direct evidence for A itself, above and beyond any evidence for (proper) subsets of A.

Möbius transforms were more generally studied by Rota [31]. Generalizations to infinite state spaces were obtained by Gilboa and Schmeidler [15], and by Marinacci [22]. The former deals with representation using only "unanimity games" but cannot guarantee countable additivity of the representation (that is, of the measure that generalizes ϕ_v above). Marinacci generalizes unanimity games to games defined on filters, and guarantees countable additivity on that space.

7.3 Representations

A capacity $v \in V$ *represents* a relation $\succsim \subset 2^Y \times 2^Y$ iff, for every $A, B \subset Y$,

$$A \succsim B \qquad \Leftrightarrow \qquad v(A) \geq v(B)$$

We mention without proof:

[26] Set functions were used in the theory of transferable utility coalitional games. In this context, Shapley [36] showed that the unanimity games form a linear basis for the space of games, and that, when a game v is described as the linear combination of the unanimity games, $\phi_v(B)$ is the coefficient of the unanimity game on B.

Remark 1 *A relation can be represented by a capacity if and only if it is a qualitative capacity.*

We first prove the following proposition, and then explain how this identifies the binary relations $\succsim_{h_t}$ that can be represented via credence functions via (1)–(2).

Proposition 4 *A relation can be represented by a belief function if and only if it is a p-monotone qualitative capacity.*

Proof. Assume that v is a belief function that represents $\succsim$. By Remark 1, $\succsim$ is a qualitative capacity. To see that it is p-monotone, let there be given $E \subset F \subset Y$, and $G \subset Y$ such that $G \cap F = \varnothing$ and $E \cup G \succ E$. Hence $v(E \cup G) > v(E)$ and thus

$$\sum_{D \subset E \cup G} \phi_v(D) > \sum_{D \subset E} \phi_v(D)$$

which means that there exists $D \subset E \cup G$, $D \nsubseteq E$ such that $\phi_v(D) > 0$. Since $D \subset E \cup G$ we also have $D \subset F \cup G$. However, $D \nsubseteq F$ because $D \cap G \neq \varnothing$ and $G \cap F = \varnothing$. This means that $v(F \cup G) > v(F)$ also holds, and $F \cup G \succ F$.

Conversely, assume that $\succsim$ is a p-monotone qualitative capacity. We wish to construct a belief function v that represents it. We will construct a totally monotone set function v that represents $\succsim$, and then (because, by the representation, we have $v(Y) > 0$), this v can be normalized to be a belief function.

The basic idea of the proof is straightforward. Assume, first, that there are no $\sim$-equivalences (apart from the trivial ones), that is, that all equivalence classes are singletons. In this case, we order the events according to $\succsim$ and assign to them weights ϕ_v inductively. Thanks to monotonicity, all subsets of a set A appear before it in the $\succsim$ ranking, and thus, when we come to set $v(A)$, the total weight assigned to its proper subsets is already given, and we only need to make sure that $\phi_v(A)$ is high enough to make $v(A)$ larger than $v(B)$ for any B that precedes A in the ranking. Notice that this proof does not rely on p-monotonicity of $\succsim$; indeed, in the absence of ties, p-monotonicity is implied by monotonicity.

In the general case, however, an $\sim$-equivalence class may contain several sets, and, importantly, also sets A, B such that $A \subset B$. This means that one

cannot assign a value to $\phi_v(A)$ without also affecting $v(B)$, which should still equal $v(A)$. As a result the argument is a little more delicate, and makes use of p-monotonicity of $\succsim$.

Let $(A_1 = \varnothing, A_2, A_3, ..., A_{2^n} = Y)$ be an enumeration of 2^Y such that $A_i \precsim A_{i+1}$ for $i < 2^n$. Let the equivalence classes of $\tilde{}$ be $(\mathcal{E}_1, ..., \mathcal{E}_K)$ with $\mathcal{E}_k = \{A_i\}_{i \in E_k}$ where $E_k = \{i_k, ..., i_{k+1} - 1\}$ (with $i_{k+1} \geq i_k + 1$). Clearly, if $A_i \in \mathcal{E}_k$ and $A_{i'} \in \mathcal{E}_{k'}$, then $A_i \succsim A_{i'}$ iff $k \geq k'$.

We will define v and ϕ_v on $\cup_{k \leq j} \mathcal{E}_k$, by induction on $j = 1, ..., K$, so that v represents $\succsim$ on $\cup_{k \leq j} \mathcal{E}_k$. Observe that, by monotonicity, if $A_i \in \mathcal{E}_k$, and $D \subset A$, then $D \in \mathcal{E}_{k'}$ for $k' \leq k$. Hence the values of v on $\cup_{k \leq j} \mathcal{E}_k$ define those of ϕ_v on $\cup_{k \leq j} \mathcal{E}_k$ and vice versa.

For $j = 1$, we define, for all $A \in \mathcal{E}_1$, $\phi_v(A) = 0$, which, by monotonicity, implies $\phi_v(D) = 0$ for every $D \subset A$ and thus also $v(A) = 0$.

Assume, then, that v and ϕ_v have been defined for $\cup_{k \leq j} \mathcal{E}_k$, so that v represents $\succsim$ on $\cup_{k \leq j} \mathcal{E}_k$, and consider $\mathcal{E}_{k+1}$. Let $\mathcal{E}_{k+1} = \{A_i\}_{i \in E_{k+1}}$. For $i \in E_{k+1}$, let

$$\alpha_i = \sum_{D \subset A_i, D \notin \mathcal{E}_{k+1}} \phi_v(D).$$

Thus, α_i is a lower bound on the value $v(A_i)$, obtained by the ϕ_v already assigned to subsets of A_i. All subsets $D \subset A_i$ such that $D \prec A_i$ are included in the summation above, but there may be other subsets $D \subset A_i$ such that $D \sim A_i$ (that is, $D \in \mathcal{E}_{k+1}$).

We will define $\phi_v(A_i) \geq 0$ so that, for all $i \in E_{k+1}$,

$$v(A_i) = M_k \equiv \max_{i \in E_{k+1}} \alpha_i + 1.$$

For a given $i \in E_{k+1}$, if A_i is a minimal set (with respect to set inclusion) in $\mathcal{E}_{k+1}$, set

$$\phi_v(A_i) = M_k - \alpha_i$$

and otherwise (i.e., if there exists $h \in E_{k+1}$, $h \neq i$, such that $A_h \subsetneq A_i$)

$$\phi_v(A_i) = 0.$$

We now wish to show that, for every $i \in E_{k+1}$, $v(A_i) = M_k$. Consider

first i such that A_i is a minimal set in $\mathcal{E}_{k+1}$. Obviously,

$$\begin{aligned} v(A_i) &= \sum_{D \subset A_i} \phi_v(D) \\ &= \sum_{D \subset A_i, D \notin \mathcal{E}_{k+1}} \phi_v(D) + \phi_v(A_i) \\ &= \alpha_i + M_k - \alpha_i \\ &= M_k. \end{aligned}$$

Next consider i such that A_i is not a minimal set in $\mathcal{E}_{k+1}$. Define

$$\mathcal{F}_i = \{ A_h \in \mathcal{E}_{k+1} \mid A_h \subset A_i \}$$

We argue that $\mathcal{F}_i$ is closed under intersection. Suppose that $B, C \in \mathcal{F}_i$. If $B \cap C \prec A_i \cap C = C$, then, by p-monotonicity and Lemma 3, $A_i \succ B$, which is known to be false because $A_i, B \in \mathcal{E}_{k+1}$ (and thus $A_i \sim B$). Hence $B \cap C \sim C$ and it follows that $B \cap C \in \mathcal{E}_{k+1}$. Clearly, $B \cap C \subset A_i$.

Let

$$B_i = \cap_{A_h \in \mathcal{F}_i} A_h \in \mathcal{F}_i.$$

(Note that B_i is a minimal set in $\mathcal{E}_{k+1}$ and thus $\phi_v(B_i)$ has been defined above, while, for all $C \in \mathcal{F}_i \setminus \{B_i\}$ we have just defined $\phi_v(C) = 0$.)

We wish to show that

$$v(A_i) = v(B_i) = M_k.$$

Clearly, every $D \subset B_i$ satisfies $D \subset A_i$. Hence $v(A_i) \geq v(B_i)$. We wish to show that for $D \subset A_i$ such that $D \nsubseteq B_i$ we have $\phi_v(D) = 0$. If $B_i \subset D$, $D \in \mathcal{F}_i$ and $\phi_v(D) = 0$ follows from the definition above. Otherwise, $B_i \nsubseteq D$ and $D \notin \mathcal{F}_i$. In particular, $D \notin \mathcal{E}_{k+1}$ and this implies that $D \in \cup_{k \leq j} \mathcal{E}_k$, and, by the induction hypothesis,

$$v(D) = \sum_{G \subset D} \phi_v(G).$$

If $\phi_v(D) > 0$, $v(D) > v(B \cap D)$ and, because v is known to represent $\succsim$ on $\cup_{k \leq j} \mathcal{E}_k$, $D \succ B \cap D$. But this would imply that $A \cap D = D \succ D \cap B$, and by p-monotonicity and Lemma 3, $A \succ B$, which is known not to be the case. We therefore conclude that – also in the case $D \notin \mathcal{E}_{k+1}$ – we have $\phi_v(D) = 0$. This concludes the proof that $v(A_i) = v(B_i) = M_k$.

It remains to be noted that the function v thus constructed represents $\succsim$ on $\cup_{k\leq j}\mathcal{E}_{k+1}$. ■

Observe that a corollary of this proposition is that a qualitative capacity (on a finite set) can be represented by a p-monotone capacity if and only if it can be represented by a belief function.

Every $Y' \subset Y$ corresponds to an event $[h_t, Y'] \in \mathcal{E}_1$. Hence, for every history h_t and p-monotone qualitative capacity, we can define a function φ_{h_t} that assigns weight to conjectures of the form $[h_t, Y']$ (and to no other conjectures), so as to ensure that φ_{h_t} induces a belief function on 2^Y that represents $\succsim_{h_t}$. Conversely, any nontrivial, σ-additive measure on $\mathcal{E}$ induces a belief function on 2^Y, and hence represents a p-monotone qualitative capacity.

The predictions $\succsim_{h_t}$ that can be represented by a credence function via (1)–(2) are precisely those for which $\succsim_{h_t}$ is a p-monotone qualitative capacity, for each h_t.

7.4 Belief Updating

Dempster [9] defined the combination of two belief functions, designed to merge information from two sources. A special case of such a combination is the situation in which one source states that a certain event, B, has occurred. This piece of information is modeled by the belief function v such that $\phi_v(B) = 1$.

It is well-known that Dempster's rule of combination in this case becomes

$$v(A|B) = \frac{v((A \cap B) \cup B^c) - v(B^c)}{1 - v(B^c)}$$

which is known as the *Dempster-Shafer updating rule* for a capacity v. Clearly, if v is additive, this rule reduces to Bayesian updating.

Gilboa and Schmeidler [13] approached the problem of updating capacities axiomatically, identifying reasonable axioms that characterize this rule.

7.5 Bayesian Beliefs

It is well-known that a capacity v is additive (that is, $v(A \cup B) = v(A) + v(B)$ whenever $A \cap B = \varnothing$) iff its Möbius transform ϕ_v is concentrated on singletons, that is, $\phi_v(A) = 0$ whenever $|A| > 1$.

Assume that v is used for decision making via maximization of Choquet integration of a non-negative utility function u, as proposed and axiomatized in Schmeidler [34]. That is, decisions are taken so as to maximize

$$\int_Y u dv = \int_0^\infty v\left(\{\, i \in Y \mid u(i) \geq t \,\}\right) dt.$$

In this case, the decision maker maximizes expected utility (specifically, satisfies Savage's axiom P2) iff v is additive.

References

[1] Hirotugu Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6(2): 127–132, 1954.

[2] Ron Alquist and Lutz Kilian. What do we learn from the price of crude oil futures? *Journal of Applied Econometrics*, 25: 539–573, 2010.

[3] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418, 1763. Communicated by Mr. Price.

[4] Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713.

[5] Rudolf Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.

[6] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27, 1967.

[7] Bruno de Finetti. Sul Significato Soggettivo della Probabilità. *Fundamenta Mathematicae*, 17: 298-329, 1931.

[8] Bruno de Finetti. La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincare*, 7(1): 1–68, 1937.

[9] Arthur. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2): 325–339, 1967.

[10] Evelyn Fix and J. L. Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical report 4, project number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[11] Evelyn Fix and J. L. Hodges. Discriminatory analysis. Nonparametric discrimination: Small sample performance. Report A193008, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.

[12] Itzhak Gilboa and Larry Samuelson. Subjectivity in inductive inference. Cowles Foundation Discussion Paper 1725, Tel Aviv University and Yale University, 2009.

[13] Itzhak Gilboa and David Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory*, 59(1): 33–49, 1993.

[14] Itzhak Gilboa and David Schmeidler. Case-based decision theory. *Quarterly Journal of Economics*, 110(3): 605–640, 1995.

[15] Itzhak Gilboa and David Schmeidler. Canonical Representation of Set Functions. *Mathmatics of Operations Research*, 20: 197–202, 1995.

[16] Itzhak Gilboa and David Schmeidler. *A Theory of Case-Based Decisions*. Cambridge University Press, Cambridge, 2001.

[17] Itzhak Gilboa and David Schmeidler. Inductive inference: An axiomatic approach. *Econometrica*, 171(1): 1–26, 2003.

[18] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

[19] David Hume. *An Enquiry Concerning Human Understanding*. Clarendon Press, Oxford, 1748.

[20] Richard Jeffrey. *Subjective Probability: The Real Thing*. Cambridge, Cambridge University Press, 2004.

[21] Dennis V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, Cambridge, 1965.

[22] Massimo Marinacci. Decomposition and Representation of Coalitional Games. *Mathmatics of Operations Research*, 21 (4): 1000–1015, 1996.

[23] Akihiko Matsui. Expected utility and case-based reasoning. *Mathematical Social Sciences*, 39(1): 1-12, 2000.

[24] John McCarthy. Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence*, 13(1–2): 27–39, 1980.

[25] Drew McDermott and John Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13(1–2): 41–72, 1980.

[26] Nils J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1): 71–87, 1986.

[27] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3): 241–288, 1986.

[28] Frank P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, pages 156–198. Harcourt, Brace and Company, New York, 1931.

[29] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1–2): 81–132, 1980.

[30] Christopher K. Riesbeck and Roger C. Schank. *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, Hilldale, New Jersey, 1989.

[31] Gian-Carlo Rota. On the foundations of combinatorial theory I: Theory of Möbius functions. *Probability Theory and Related Fields*, 2(4), 340–368, 1964.

[32] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, New York, 1972 (originally 1954).

[33] Roger C. Schank. *Explanation Patterns: Understanding Mechanically and Creatively*. Lawrence Erlbaum Associates, Hilldale, New Jersey, 1986.

[34] David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3): 571–587, 1989.

[35] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.

[36] Lloyd S. Shapley. A Value for n-Person Games, 1953.

[37] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London and New York, 1986.

[38] Ray J. Solomonoff. A formal theory of inductive inference I,II. *Information Control*, 7(1,2): 1–22, 224–254, 1964.

[39] Francis Voorbraak. On the justification of Dempster's rule of combination. *Artificial Intelligence*, 48: 171–197, 1991.

[40] Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, London, 1922.