

Dynamics of Inductive Inference in a Unified Framework¹

Itzhak Gilboa,² Larry Samuelson,³ David Schmeidler⁴

September 12, 2011

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 event 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.

¹We thank Daron Acemoglu, Dirk Bergemann, Eddie Dekel, Drew Fudenberg, Gabi Gayer, Offer Lieberman, and George Mailath for comments and suggestions.

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

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

⁴The Ohio State University and Tel-Aviv University.

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

September 12, 2011

Contents

1	Introduction	1
2	The Framework	4
2.1	The Environment	4
2.2	Predictions	6
2.3	Updating	8
3	Special Cases	11
3.1	Bayesian Reasoning	11
3.2	Case-Based Reasoning	13
3.3	Rule-Based Reasoning	16
3.4	Combined Models	17
3.5	How Would We Know and Why Would We Care?	18
4	Dynamics of Reasoning Methods	19
4.1	When is Bayesian Reasoning Fragile?	19
4.1.1	Assumptions	19
4.1.2	Result	22
4.1.3	Weights of Credence	23
4.1.4	The <i>iid</i> Case	25
4.2	When will Bayesianism Prevail?	29
4.3	Probabilistic Reasoning	32
5	Concluding Remarks	35
5.1	Methods for Generating Conjectures	35
5.2	Probabilistic Conjectures	36
5.3	Single-Conjecture Predictions	36
5.4	Decision Theory	37
6	Appendix: Proofs	38
6.1	Proof of Proposition 1	38
6.2	Proof of Proposition 2	38

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

September 12, 2011

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 events 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

¹In 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 events 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's 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.

²Indeed, Alquist and Kilian (2010) 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 events 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. 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.

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 $t \in \{0, 1, \dots\}$ 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

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

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., $\omega : \{0, 1, \dots\} \rightarrow X \times Y$. We let $(\omega_X(t), \omega_Y(t))$ denote the element (x_t, y_t) 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

$$h_t(\omega) = (\omega(0), \dots, \omega(t-1), \omega_X(t))$$

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 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 , i.e., $H_t = \{h_t(\omega) \mid \omega \in \Omega\}$. For a history h_t we define the corresponding event

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

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 the next outcome being 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 $A = \{\omega\}$, and such conjectures will suffice to capture Bayesian reasoning. 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 $A \cap [h_t] \neq \emptyset$. The set of conjectures that have not been refuted by history h_t and predict an outcome in Y' is⁵

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

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

⁴Note 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 $\omega \in \Omega$.

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

this comparison by using a “credence function” φ_{h_t} . Formally, φ_{h_t} is a finite, σ -additive measure on the 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 The agent ranks Y' as “at least as likely as” Y'' , denoted $Y' \succeq_{h_t} Y''$, iff

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

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.

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 $\cup_{Y' \subseteq Y} \mathcal{A}(h_t, Y')$, 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

⁶There 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.

⁷The weighting function φ_{h_t} is equivalent to a belief function in the Dempster-Shafer theory of evidence (Dempster [10], Shafer [38]).

⁸Recall 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.

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

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	$\{\{\omega\} \mid \omega \in A\}$ 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 single-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. Section 5.2 sketches the beginnings of a generalization to non-deterministic conjectures.

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,

¹⁰In 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 $A \cap [h_t] \neq \emptyset$.

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 φ_\emptyset of related conjectures to zero. However, given a history h_t that includes this event, conjectures that involve similar events 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 competing 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 ($|Y| = 2$) and that conjecture A makes predictions at each of the periods $t = 0, \dots, 100$, 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.” 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 φ_\emptyset , perhaps by assigning weights $\varphi_\emptyset(A) = 100\varphi_\emptyset(A')$, and can then simply use φ_\emptyset 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 .¹¹ 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 φ_\emptyset 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 $\varphi_\emptyset(A)$ and $\varphi_\emptyset(A')$ 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 $\{\succ_{h_t} \subset 2^Y \times 2^Y\}_{t \geq 0, h_t \in H_t}$. 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 (Ω, \mathcal{A}) . Then there exists a measure φ on (Ω, \mathcal{A}) such that, at each h_t and for every $Y', Y'' \subset Y$,*

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

¹¹Alternatively, if A predicts incorrectly during some of the first 100 periods, it will subsequently be excluded and hence this choice of φ_\emptyset will not interfere with further predictions.

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 [10], Shafer [38]) updating of the belief function defined by φ , in face of the evidence $[h_t]$. This updating rule has been axiomatized by Gilboa and Schmeidler [16] in the context of Choquet expected utility maximization.¹²

3 Special Cases

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

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.¹³ 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’s rule.

¹²This updating is a special case of Dempster’s rule of combination, in which the belief function defined by φ is combined with the belief function that attaches weight 1 to the events that contain the conjecture $[h_t]$ (and zero to all other events). This special case of Dempster’s rule of combination does not suffer from common criticisms of the Dempster-Shafer theory, such as those leveled by Voorbraak [41].

¹³Bayesian 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 [26]). Relying on the axiomatic approach of Ramsey [32], de Finetti [8, 9], and Savage [35], 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 [31]). Within the philosophy of science, notable proponents of the Bayesian approach include Carnap [5] and Jeffrey [22].

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

$$\mathcal{B} = \{\{\omega\} \mid \omega \in \Omega\} \subset \mathcal{A}. \quad (3)$$

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

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

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:

Lemma 1 *Let p be a probability measure on (Ω, \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 $Y' \subset Y$*

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

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. In addition,

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

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 $A \in \mathcal{A} \setminus \mathcal{B} > 0$) 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} , will not satisfy de Finetti’s [8, 9] cancellation axiom: it can be the case that, for two events, B, C , $B \succsim_{h_t} C$ but not $B \setminus C \succsim_{h_t} C \setminus B$. In addition, if we use the weight function to make decisions by maximization of the Choquet integral of a utility function, the maximization will fail to satisfy Savage’s [35] “sure-thing principle” (axiom P2). As a result, upon adding decisions to our model of beliefs (cf. Section 5.4), we have a converse to Lemma 1: the decision maker will be Bayesian if and *only if* (4) holds.

3.2 Case-Based Reasoning

Case-based reasoning is also a special case of our model.¹⁴

We first introduce a simple model of case-based reasoning in which case-based prediction is equivalent to kernel classification.¹⁵ The agent has a

¹⁴Analogical reasoning was explicitly discussed by Hume [21], and received attention in the twentieth century in the guise of case-based reasoning (Riesbeck and Schank [34], Schank [36]), leading to the formal models and axiomatizations of Gilboa and Schmeidler [17, 18, 19].

¹⁵See Akaike [1] and Silverman [39].

similarity function over the characteristics,

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

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' .

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, $\beta \rightarrow 0$, 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 case-based reasoning:

Lemma 2 *Let there be given $s : X \times X \rightarrow \mathbb{R}_+$ and $\beta \leq 1$. There exists a model 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

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

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

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

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

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

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 $\{A_{i,t,\omega_X(i),\omega_X(t)} \mid i < t\}$ 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 $i = 0, 1, \dots, t-1$). 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. 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 $(i - k + 1), \dots, i$ are identical to observations $(t - k + 1), \dots, t$, and so forth.

3.3 Rule-Based Reasoning

The model can accommodate many other forms of reasoning, often referred to as “rule-based reasoning.”¹⁶ 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

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

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

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

(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.”¹⁷

¹⁶We 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 [28], McDermott and Doyle [29], Reiter [33]), probabilistic (Nilsson [30]), and a variety of other new logics.

¹⁷Holland’s [20] genetic algorithms address classification problem where the value of y is to be determined by the values of $x = (x^1, \dots, x^m)$, 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

An association rule will be excluded from the summation defining $\varphi(\mathcal{A}(h_t))$ 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

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

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 models 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

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

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.

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

If ε is small, the non-Bayesian conjectures will play a relatively minor role in determining predictions, as long as history proceeds along a path that had a high a-priori probability. However, suppose that the reasoner faces a surprising event, such as the September 11 attacks or the Lehman Brothers' collapse. If the agent had assigned the event zero probability, Bayesian updating will not be well-defined. In this case, the non-Bayesian conjectures may 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 event in question had a nonzero but very small prior probability, non-Bayesian reasoning will again be relatively more important. Once the event has occurred, conditional probabilities are 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. When a small probability event occurs, it is as if the agent tells herself, "I do have my updated Bayesian beliefs, but I start doubting my probability assessments; after all, according to these very same assessments, it used to be very unlikely to find ourselves where we are. Hence, it might be a good idea to consider other modes of reasoning as well."

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.¹⁸ 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.¹⁹ 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.

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:

¹⁸Relatedly, Matsui [27] demonstrated that expected utility maximization and case-based decision theory lead to equivalent sets of feasible outcomes.

¹⁹Naturally, such a task requires additional assumptions on the structure of the prior probability.

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:

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

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 event cannot be too much smaller than that assigned to another event at the same period t . Thus, one may assume that there exists $M > 1$ such that, for every t and every $h_t, h'_t \in H_t$,

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

We weaken this condition still further, allowing M to depend on t , and assume only that the ratio between the probabilities of two events 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 $p > 0.5$.²⁰ 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

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

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

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

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., 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, (1) for every i, i', t, t', x, x' and z, z' with $t - i = t' - i'$, and $t' < t$,*

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

²⁰For 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 .

and (2) for every $t, x, z \in X$ and $i < i' < t$,

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

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 :

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

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

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

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 model 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.²¹

²¹At 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 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.

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 is 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 $h_t = ((x_0, y_0), \dots, (x_{t-1}, y_{t-1}), x_t)$, 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 governed by such a measure:

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

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

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 .²²

We can now state

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

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

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.

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

²²Observe 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.

every t and every h_t ,

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

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 relative 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. In between these two possibilities, if the 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), then the evolution of her reasoning will depend on the rate at which the $\varphi(\{A_{i,t,x,z}\})$ decline and the nature of her Bayesian prior. To illustrate the latter, 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. The agent will thus remain Bayesian if her Bayesian prior contains the correct data generating process, but will otherwise slip into case-based reasoning. ■

Example 3 This example illustrates another case in which not all weight is attached to Bayesian conjectures, so it is not obvious that Bayesian reasoning will survive, but Assumption 3 does not hold, so it is also not obvious that Bayesian reasoning will disappear.

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.

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 $y = 0$ or $y = 1$ 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 least 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 hypothesis 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 4 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 5 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

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

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.²³ Once again, Assumption 2 fails. The remaining weight, ε , is split among the case-based conjectures.

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 $\{\omega^{k'}\}$ for $k' \neq k$ are refuted. In contrast, the conjecture $\{\omega^k\}$ 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

²³Observe 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.

of reasoning remaining the dominant one (for large t). Clearly, this will only be true at the states $\{\omega^k\}$. 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 6 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 $1, 1, \dots, 1$, 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 $1, 1, \dots, 1$, 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 is 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 to 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,

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

(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 $k \ll t$. 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, as $t' > t$ grows, a repetition of the arguments behind Proposition ?? ensures that 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

$$0, 0, \dots, 0, 1, 0, 0, \dots, 0, 1, \dots$$

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 Probabilistic Conjectures

An important next step is to extend this framework to probabilistic conjectures. Conjectures would then be represented by probability distributions rather than by sets of states. The Bayesian conjectures in such an extension are straightforward, and consist of probability distributions over states. Each such distribution f has an a priori weight $\varphi(\{f\})$. If the support of φ is contained within the set of Bayesian conjectures, then φ is simply the Bayesian prior. Given a history h_t , the conjecture f is no longer classified dichotomously into “consistent with h_t ” or “inconsistent with h_t .” Rather, it is continuously ranked in $[0, 1]$ according to the probability of history h_t given theory f , that is, according to the theory’s likelihood function at h_t . Multiplying the likelihood function by the a-priori weight $\varphi(\{f\})$ leads to a natural measure of the belief in theory f following history h_t . Indeed, this is, up to renormalization, precisely the result of a Bayesian update over the Bayesian conjectures.

The specification of non-Bayesian conjectures is less clear. Should these be formulated as sets of distributions over states, or as distributions over sets of states, some combination of these generalizations, or something else? Finding such an appropriate generalization is a topic for further research.

5.3 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

[42] definition of induction.²⁴ If, by contrast, we apply this method to case-based conjectures, we end up with nearest-neighbor approaches (see Cover and Hart [7] and Fix and Hodges [11, 12]) rather than with the case-based aggregation discussed here.

5.4 Decision Theory

In order to explore the implications of our analysis to decision making, the framework needs to incorporate acts and payoffs, and to specify the interaction between the agent's choices and the underlying process. There are situations in which this interaction is practically non-existent. For example, a small trader in the stock market may assume that her actions have no effect on future prices. In this case, the decision problem is in close relationship to a prediction problem: the payoff at each period is a function of the quality of the guess made, but no additional complications arise. Other examples of this type include a physician who has to make diagnoses or treatment decisions for a different patient each period, or a graduate admissions officer who has to make admission decisions for consecutive candidates.

However, many choice situations require one to go beyond predictions, and to consider the effect that one's choices might have on the unfolding of the process in the future. In these situations, it is conceptually simplest to assume that the agent makes one choice of an act (or a strategy) at the outset, then history unfolds, nature determines the state of the world, and the agent's utility is determined by the resulting outcome. In this case, each act f associates outcomes with states ω as in a standard Savage model. But our framework needs to be augmented before it can be used to generate beliefs over this state space. The reason is that many conjectures in the framework – such as case-based conjectures, or those corresponding to association rules – only constrain the values of y_t given x_t , but remain silent on the evolution of the x_t 's in the future. Such conjectures are all one needs to make conditional predictions at a specific period t , but if one engages in long-run predictions, one has to ask oneself not only which y_τ are likely to occur given x_τ for $\tau > t$, but also which x_τ are likely to be observed in the future.

²⁴See Solomonoff [40], who suggested to couple this preference for simplicity with Kolmogorov complexity measure to yield a theory of philosophy of science. Gilboa and Samuelson [15] discuss the optimal selection of the preference relation over theories in this context.

6 Appendix: Proofs

6.1 Proof of Proposition 1

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

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

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 and non-tautological 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 event 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

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

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,

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

Then,

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

Choose T large enough so that

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

From now on, assume that $t \geq T$.

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

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

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

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

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

$$\sum_{\tau=1}^T S_\tau < \xi [Q(t)]^2 |X|^2 \binom{t}{2}$$

and, using (17a) we obtain

$$\frac{\varphi(\mathcal{CB})}{2} < \sum_{\tau=1}^T S_{\tau} < \xi [Q(t)]^2 |X|^2 \binom{t}{2}.$$

Define

$$R(t) = 2 [Q(t)]^2 |X|^2 \binom{t}{2}$$

and observe that it is a polynomial in t .

Thus, we have

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

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

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

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

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

Combining (16) and (18) we obtain

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

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

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] Gustave Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5 (Grenoble): 131–295, 1953–54.
- [7] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27, 1967.
- [8] Bruno de Finetti. Sul Significato Soggettivo della Probabilità. *Fundamenta Mathematicae*, 17: 298–329, 1931.
- [9] Bruno de Finetti. La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincare*, 7(1): 1–68, 1937.
- [10] Arthur. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2): 325–339, 1967.
- [11] 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.
- [12] Evelyn Fix and J. L. Hodges. Discriminatory analysis. Nonparametric discrimination: Small sample performance. Report A193008, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.
- [13] Peter Gärdenfors. Induction, conceptual spaces and AI. *Philosophy of Science*, 57(1): 78–95, 1990.
- [14] Itzhak Gilboa. *Theory of Decision under Uncertainty*. Cambridge University Press, Cambridge, 2009.

- [15] Itzhak Gilboa and Larry Samuelson. Subjectivity in inductive inference. Cowles Foundation Discussion Paper 1725, Tel Aviv University and Yale University, 2009.
- [16] Itzhak Gilboa and David Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory*, 59(1): 33–49, 1993.
- [17] Itzhak Gilboa and David Schmeidler. Case-based decision theory. *Quarterly Journal of Economics*, 110(3): 605–640, 1995.
- [18] Itzhak Gilboa and David Schmeidler. *A Theory of Case-Based Decisions*. Cambridge University Press, Cambridge, 2001.
- [19] Itzhak Gilboa and David Schmeidler. Inductive inference: An axiomatic approach. *Econometrica*, 171(1): 1–26, 2003.
- [20] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.
- [21] David Hume. *An Enquiry Concerning Human Understanding*. Clarendon Press, Oxford, 1748.
- [22] Richard Jeffrey. *Subjective Probability: The Real Thing*. Cambridge, Cambridge University Press, 2004.
- [23] Frank H. Knight. *Risk, Uncertainty, and Profit*. Boston, New York: Houghton Mifflin, 1921.
- [24] David M. Kreps. *Notes on the Theory of Choice*. Westview Press, Boulder, Colorado, 1988.
- [25] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.
- [26] Dennis V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, Cambridge, 1965.
- [27] Akihiko Matsui. Expected utility and case-based reasoning. *Mathematical Social Sciences*, 39(1): 1–12, 2000.
- [28] John McCarthy. Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence*, 13(1–2): 27–39, 1980.

- [29] Drew McDermott and John Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13(1–2): 41–72, 1980.
- [30] Nils J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1): 71–87, 1986.
- [31] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3): 241–288, 1986.
- [32] 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.
- [33] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1–2): 81–132, 1980.
- [34] Christopher K. Riesbeck and Roger C. Schank. *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, Hilldale, New Jersey, 1989.
- [35] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, New York, 1972 (originally 1954).
- [36] Roger C. Schank. *Explanation Patterns: Understanding Mechanically and Creatively*. Lawrence Erlbaum Associates, Hilldale, New Jersey, 1986.
- [37] David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3): 571–587, 1989.
- [38] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.
- [39] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London and New York, 1986.
- [40] Ray J. Solomonoff. A formal theory of inductive inference I,II. *Information Control*, 7(1,2): 1–22, 224–254, 1964.
- [41] Francis Voorbraak. On the justification of Dempster’s rule of combination. *Artificial Intelligence*, 48: 171–197, 1991.
- [42] Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, London, 1922.