DYNAMICS OF INDUCTIVE INFERENCE IN A UNIFIED FRAMEWORK

By

Itzhak Gilboa, Larry Samuelson and David Schmeidler

July 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1811

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Dynamics of Inductive Inference in a Unified Framework[1]

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

February 4, 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, positively or normatively, how the various modes of inductive inference can be combined and how their relative weights change endogenously. 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 show that even random data can make certain theories seem plausible and hence increase the weight of rule-based vs. case-based reasoning, leading the agent in some cases to cycle between being rule-based and case-based. We identify conditions under which minmax regret criteria will not be effective.

[1]We thank Dirk Bergemann, Eddie Dekel, Drew Fudernberg, Gabi Gayer, Offer Lieberman, George Mailath, the editors and three referees for comments and suggestions.

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

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

[4]The Ohio State University and Tel-Aviv University.

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

February 4, 2011

Contents

Dynamics of Inductive Inference in a Unified Framework

Itzhak Gilboa, Larry Samuelson, David Schmeidler

February 4, 2011

1 Introduction

How should we model an agent who learns and updates her beliefs? Learning and inductive inference are extensively studied in a variety of fields, ranging from statistics and machine learning to psychology and artificial intelligence. It makes sense to assume that, when statistical analysis is possible, as in the case of many observations of iid random variables, rational agents will perform such analysis more or less correctly in the long run. By contrast, our interest is in the way economists model agents who face problems that do not naturally lend themselves to statistical analysis. For example, when predicting financial crises and economic growth, the eruptions of wars and revolutions, or the outcome of elections, an agent surely relies on statistical learning, but also has to use other methods to reason about economic, political, and social trends. It is this type of reasoning that is the focus on this paper.

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? We wish to model three types of reasoning.[1]

[1] In personal conversation, a hedge fund principal indicated that his fund used all three

The mode of reasoning most widely used in 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 in order to make new predictions.

An alternative mode of reasoning is *case-based*. This means that the agent considers past observations and predicts an outcome that appeared more often in those past cases that are considered similar or otherwise relevant. If all past observations are considered equally similar (and relevant), case-based prediction is simply the mode, that is, the outcome that is most frequent in the database. If, by contrast, the agent uses a similarity function that puts all its weight on the most recent outcome, her prediction will simply be that outcome.[2] However, more interesting similarity functions can use more data for the judgment of relevance of past cases. For example, the agent may argue that the current state of conflict in the Middle East is reminiscent of the state of affairs in 1991 or in 2003, and hence 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. Case-based and rule-based agents can update the probabilities they attach to the validity of various analogies or rules in light of their experience, much as would a Bayesian. A Bayesian will base her prior distribution

methods of reasoning introduced in this section in predicting the likelihood of mortgage defaults.

[2]Indeed, Alquist and Kilian (2010) find that the best prediction of the future price of oil is the current price.

on analogies to similar past cases, as well as on general rules that she has observed. A precise definition of rule-based reasoning is especially elusive, leaving us instead with an "I know it when I see it" approach. To make one boundary precise, we say that reasoning is "Bayesian" if all past analogies, regularities, and other prior information can be summarized in a prior probability distribution over the possible remaining histories, with all subsequent reasoning captured by standard Bayesian updating.[3]

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, which captures a way of thinking about how events in the world will develop, while 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.

Within this framework, one may pose both positive and normative questions, having to do with the conjectures that the agent considers, and the weight she attaches to them. An example of a positive question would be, what are the weights that people tend to attach to various conjectures? Or, given certain initial weights, how do they evolve, and what would be the long-term relative weights of different types of reasoning? Section 4 addresses such questions. Alternatively, one may ask more normative questions, such as which conjectures should be considered and what is an appropriate way to assign weights to them, so as to lead to effective learning. Section 5 deals with such questions.

Specifically, in Section 4.1 we consider the long-term weight of Bayesian reasoning relative to other modes of reasoning. We suggest conditions under which Bayesian reasoning will give way to other modes of reasoning, and

[3]The belief that all uncertainty can be quantified by (possibly subjective) probabilities is at the heart of the Bayesian approach. It is this belief that has been challenged by non-Bayesian thinkers such as Knight [23].

alternative conditions under which the opposite conclusion holds. We discuss these conditions in an attempt to distinguish between situations in which the Bayesian approach is likely to be robust and situations in which it is not. Section 4.2 deals with a similar question, studying the evolution of relative weights of case-based and rule-based reasoning. Here, we show that even random data can occasionally give rise to belief in a specific theory and hence in rule-based reasoning in general, until the theory is refuted and agents resort to case-based reasoning, potentially leading them to cycle between case-based and rule-based reasoning. Section 5 considers a normative question. We assume that the agent can choose the weight assigned to conjectures, for instance by putting more weight on conjectures that did well, or that would have minimized regret in the past. We show that in the absence of additional knowledge about the nature of the problem, these criteria do not lead to effective predictions.

2 The Framework

2.1 The Environment

At each period $t \in \{0, 1, \ldots\}$ there is a *characteristic* $x_t \in X$ and an *outcome* $y_t \in Y$. The sets X and Y are assumed to be finite and non-empty, with Y containing at least two possible outcomes.[4]

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, as well as 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. Our preferred interpretation is that this lack of structure reflects the agent's lack of knowledge about the data generating process—we are most interested in cases in which the agent has no certain knowledge that she can bring to bear on the prediction problem.

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

For example, we do not think of statistical inference, in which the agent knows she faces a sequence of independent random variables from a fixed distribution, as our prime application. This is in keeping with our example of an agent who must predict long-term movements in the price of oil, rather than daily fluctuations of the price around a long-term trend.[5]

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

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

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

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

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

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

consisting of all states that are compatible with the history h_t and with 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

[5]When statistical learning is possible, we would be most interested in the unstructured learning process that remains after the agent has learned what she can from such inference.

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. $\succsim_{h_t}$ is assumed to be a weak order that is monotone with respect to set inclusion.

In our example, $Y = \{0, 1\}$, and the only interesting subsets to compare are those consisting of specific outcomes, $\{0\}$ and $\{1\}$. In a richer model Y could consist of all possible prices of oil, and we would allow the agent to consider subsets of Y of the form "the price of oil will exceed \$100 per barrel" or "the price of oil will be below \$80 a barrel," and to rank some such subsets as being more likely than others. Hence, the agent may view a price of oil above \$100 as being more likely than a price under \$100, which is in turn more likely than a price of precisely \$110; or, she may view an increase in price as more likely than a decrease, and so forth.

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.

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

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 previous observations and that predict an outcome in Y'. She then considers the weight of credence attached to this set of conjectures. The agent considers the set of outcomes Y' as more likely than the set Y'' if and only if the former attains a higher weight of credence than the latter.

Formally, suppose that the agent has observed history h_t in period t and considers the set of outcomes Y'. A conjecture $A \in \mathcal{A}$ has not been refuted by history h_t if $A \cap [h_t] \neq \varnothing$. The set of conjectures that have not been

refuted by history h_t *and* predict an outcome in Y' is[6]

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

The agent evaluates the relative likelihoods of outcomes Y' and Y'', at history h_t, by comparing the sets $\mathcal{A}(h_t, Y')$ and $\mathcal{A}(h_t, Y'')$. The agent makes this comparison by using a "weight of credence" function φ_{h_t}. Formally, φ_{h_t} is a finite, σ-additive measure on sigma-algebra $\mathcal{E} \subset 2^{\mathcal{A}}$ to be defined shortly.[7] 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''.[8] The agent ranks Y' as "at least as likely as" Y'', denoted $Y' \succsim_{h_t} Y''$, iff

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

The prediction rule given by (2) exhibits a standard feature of information updating. Refuted conjectures are excluded from every set of the form $\mathcal{A}(h_t, Y')$ and hence (cf. (1)) are excluded from the prediction process at h_t. 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.[9] 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

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

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

[8]The weighting function φ_{h_t} is equivalent to a belief function in the Dempster-Shafer theory of evidence (Dempster [10], Shafer [37]).

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

only a subset of conjectures, $\cup_{Y' \subsetneq Y} \mathcal{A}(h_t, Y')$, is relevant for prediction at h_t, and the definition of φ_{h_t} outside this set is irrelevant. Over this set, the weighting function φ_{h_t} can still depend on the history h_t. We interpret $\varphi_{\varnothing}$ as the model the agent uses at the outset of the prediction problem. As the evidence unfolds, the agent may reevaluate the relative likelihoods of some conjectures, allowing φ_{h_t} to differ from $\varphi_{\varnothing}$.

A first step is obvious. Since predictions will be made by comparing the φ_{h_t} value of subsets of the type $\mathcal{A}(h_t, Y')$, we need to make sure that these are measurable. Let $\mathcal{E}_0$ be the σ-algebra generated by all such sets,

$$\{\mathcal{A}(h_t, Y')\}_{t \geq 0, h_t \in H_t, Y' \subset Y}.$$

For every measurable event, $A \in \mathcal{A}$, 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 $\mathcal{E}_1$ be the σ-algebra generated by all such sets for $A \in \mathcal{A}$. Similarly, the set of singletons in a conjecture will also be of interest, and we let $\mathcal{E}_2$ be the σ-algebra generated by all sets,

$$\{\{\omega\} \,|\omega \in A\}$$

for $A \in \mathcal{A}$.[10] Finally, we define $\mathcal{E}$ as the σ-algebra that is generated by $\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2$ and define a *model* φ_{h_t} to be a (σ-additive) probability measure on $\mathcal{E}$.

The use of 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.[11] To confirm this, Section 3.1 shows that if we restrict attention to singe-state conjectures, then we have the familiar framework of Bayesian reasoning. Expanding the framework to encompass multi-state conjectures is necessary if we are to capture case-based and rule-based reasoning (cf. Sections 3.2 and 3.3).

[10] The collection $\mathcal{E}_1$ 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.

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

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 6.2 sketches the beginnings of a generalization to non-deterministic conjectures.

It will be useful to have notation for the set of conjectures, *in a subset* $\mathcal{D} \in \mathcal{E}$, that are relevant for prediction at history h_t:

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

Notice that

$$\mathcal{D}(h_t) = \cup_{Y' \subsetneq Y} (\ \mathcal{D} \cap \mathcal{A}(h_t, Y')).$$

Hence, $\mathcal{D}(h_t) \in \mathcal{E}$. It is the set of conjectures in $\mathcal{D}$ that have not been refuted by h_t and that could lend their weight to *some* non-tautological ($Y' \subsetneq Y$) prediction after history h_t, and $\varphi_{h_t}(\mathcal{D}(h_t))$ is be the total weight of credence for these conjectures.[12]

2.3 Updating

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

Our generalized model allows a second avenue for learning—the weighting function φ_{h_t} is allowed to 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

[12]The restriction of attention to non-tautological conjectures is introduced in order to render the numerical values of $\varphi(\mathcal{D}(h_t))$, for different classes $\mathcal{D}$, more intuitive. Our main result also holds, and is in fact easier to prove, if one leaves in the set $\mathcal{D}(h_t)$ also conjectures that are unrefuted and that do not restrict the prediction $y \in Y$ in any way.

retain their original relative weights, as well as with the notion of a likelihood function, which can only decrease in value as data are gathered.

We can obviously expect φ_{h_t} to vary with h_t if the agent is initially unaware of some conjectures. Such a conjecture will be assigned a zero weight at the outset, but a positive weight at a history h_t that brings the conjecture to mind. For example, it is possible that prior to September 11, 2001 the agent had not imagined that terrorists might fly commercial airliners into buildings. This unawareness is naturally captured by setting $\varphi_\varnothing$ of related conjectures to zero. However, given a history h_t that includes this 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 $t = 0, ..., 100$, 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.

However, rewarding conjectures (or experts) for passing more prediction tests does not require that φ_{h_t} depend on h_t. Instead, these rewards can be built into a function φ that is independent of h_t. In the example above, at time $t = 0$ the agent already knows that conjecture A' will be irrelevant for the first 100 observations, and will join the game only at period $t = 100$. The agent can then build this comparison into the function $\varphi_\varnothing$, perhaps by assigning weights $\varphi_\varnothing(A) = 100\varphi_\varnothing(A')$, and can then simply use $\varphi_\varnothing$ throughout. Thus, if at time $t = 100$ conjecture A is still in the game, it will have a much higher weight than would A'.[13] There is then no need to alter φ

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

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

The more general point is that $\{\varphi_{h_t}\}_{t\geq 0, h_t \in H_t}$ is under-identified by the rankings $\left\{\succsim_{h_t} \subset 2^Y \times 2^Y\right\}_{t\geq 0, h_t \in H_t}$. Many different models $\{\varphi_{h_t}\}_{t\geq 0, h_t \in H_t}$ 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 $(\Omega, \mathcal{A}\}$. Then there exists a measure φ on $(\Omega, \mathcal{A}\}$ such that, at each h_t and for every $Y', Y'' \subset Y$,*

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

It thus sacrifices no generality to work with a function φ that is unchanged as history unfolds. Whether one wants to respond to the unfolding data by incorporating conjectures of which the agent was initially unaware or by increasing the weight of conjectures that have passed many tests, the resulting system of weight functions $\left(\varphi_{h_t}\right)_{h_t}$ is equivalent to a single function

dictions.

φ that does not change with history. We accordingly hereafter drop the h_t subscript on φ and work with an unchanging φ.

Observe that, when φ is independent of history, the updating rule inherent in (1)–(2) is equivalent to the Dempster-Shafer (cf. Dempster [10], Shafer [37]) 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.[14]

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 appeared explicitly in the writings of Bayes [3].[15] 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 [31], de Finetti [8, 9], and Savage [34], 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 [30]). Within the philosophy of science, notable proponents of the Bayesian approach include Carnap [5] and Jeffrey [22]. These 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.

[14] This 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 [40].

[15] Precursors can be found in the early days of probability; see Bernoulli [4].

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

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

Notice that $\mathcal{B}$ is an element of $\mathcal{E}$. Moreover, for every history h_t, $\mathcal{B}(h_t)$ is given by

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

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

Each of the Bayesian conjectures thus fully specifies a single state of the world. In our price-of-oil example, a specific scenario might be that, at each t, demand for oil will increase, the level of conflict will not, and the price of oil will increase. This identifies a unique state ω with $\omega_X(t) = (1, 0)$ and $\omega_Y(t) = 1$ for all t, and the corresponding conjecture is $A = \{\omega\}$.

A Bayesian agent will attach credence to no other conjectures, i.e.,

$$\varphi(\{A \in \mathcal{A} \mid |A| > 1\}) = 0.$$

We can now state:

Observation 1 *Let p be a probability on $(\Omega, \mathcal{A})$. There exists a model φ_p such that $\varphi(\mathcal{A} \backslash \mathcal{B}) = 0$ and such that for every history h_t, there is a constant $\lambda > 0$ for which, for every $Y' \subset Y$*

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

This observation is verified by constructing the model $\varphi_p(\{\{\omega\} \mid \omega \in A\}) = p(A)$ for $A \in \mathcal{A}$, attaching to each set of singleton conjectures a weight of credence equal to the prior probability of the corresponding event. It is easy to verify that φ_p satisfies the equality of Observation 1 and that it is the unique such model, up to multiplication by a positive constant.

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.

Given that our model captures Bayesian reasoning via an assumption that conjectures contain only a single state each, it is worth noting that an agent

who assigns positive weight to non-Bayesian conjectures (i.e., $\varphi(\mathcal{A}\backslash\mathcal{B}) > 0$) will not be "Bayesian" by any common definition of the term. For example, suppose that $A = \{\omega_1, \omega_2\}$ and $\varphi(\{A\}) = \delta > 0$. Such an agent can be viewed as arguing, "I think that one of ω_1 or ω_2 might occur, and I put a weight $\delta > 0$ on this conjecture, but I cannot divide this weight between the two states." Intuitively, this abandons the Bayesian tenet of quantifying all uncertainty in terms of probabilities. Formally, the corresponding rankings of subsets of outcomes, $\succsim_{h_t}$, 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 $B\backslash C \succsim_{h_t} C\backslash B$. In addition, if we use the weight function to make decisions by maximization of a Choquet integral of a utility function, the maximization will fail to satisfy Savage's [34] "sure-thing principle"(axiom P2).[16] As a result, especially upon adding decisions to our model of beliefs (cf. Section 6.4), we have a converse to Observation 1: the decision maker will be Bayesian if and *only if* $\varphi(\mathcal{A}\backslash\mathcal{B}) = 0$.

3.2 Case-Based Reasoning

Analogical reasoning was explicitly discussed by Hume [21], and received attention in the twentieth century in the guise of case-based reasoning (Riesbeck and Schank [33], Schank [35]), leading to the formal models and axiomatizations of Gilboa and Schmeidler [17, 18, 19].

We consider here a very simple version in which case-based prediction is equivalent to kernel classification.[17] The agent has a similarity function over the characteristics,

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

and a memory decay factor $\beta \leq 1$. Given history $h_t = h_t(\omega) \in H_t$, a possible outcome $y \in Y$ is assigned the weight

$$S(h_t, y) = \sum_{i=0}^{t-1} \beta^{t-i} s(\omega_X(i), \omega_X(t)) \mathbf{1}_{\{\omega_Y(i)=y\}},$$

[16] If positive weight is assigned to non-Bayesian conjectures, one should specify how expected utility maximization is generalized to a theory of decision making where beliefs are given by a function φ that is not generally additive. A well-known such generalization is the maximization of a Choquet [6] integral suggested by Schmeidler [36]. See Gilboa [14] for details and precise definitions. The axiomatic systems of de Finetti and Savage are also given in Kreps [24].

[17] See Akaike [1] and Silverman [38].

where **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 the outcome 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 the 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, $\beta \to 0$, the maximizer of $S(h_t, \cdot)$ will be the most recent observation, $\omega_Y(t-1)$. Thus, when the similarity function is constant, case-based reasoning can be viewed as a simultaneous (and smooth) generalization of prediction by empirical frequencies on the one hand, and of prediction by recency on the other hand. Clearly, more interesting generalizations are possible when the similarity function isn't constant, and uses the information given in X to make more informed judgments.

To embed case-based reasoning in our framework, we first define case-based conjectures as follows. 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 in period i are given by x and are given in period t by z, *then* periods i and t will produce the same outcome (value of y). Notice that in contrast to the Bayesian conjectures, a single case-based conjecture consists of many states: $A_{i,t,x,z}$ does not restrict the values of $\omega_X(k)$ or $\omega_Y(k)$ for $k \neq i, t$.

Let the set of all conjectures of this type be denoted by

$$\mathcal{CB} = \{A_{i,t,x,z} \mid i < t \leq T, x, z \in X\} \subset \mathcal{A}. \tag{4}$$

For example, our oil-price predictor may focus only on the years in which demand and conflict had the same trends as in the current period, and make her prediction based on the prevalence of price increases in these periods. This would correspond to the similarity function

$$s\left(\left(x^1, x^2\right), \left(z^1, z^2\right)\right) = \begin{cases} 1 & x^1 = z^1, x^2 = z^2 \\ 0 & otherwise \end{cases}.$$

Alternatively, the agent may assign some weight also to past periods that resembled the current period only in one aspect, and use a similarity function such as

$$s\left(\left(x^1,x^2\right),\left(z^1,z^2\right)\right)=\begin{cases}1 & x^1=z^1,x^2=z^2\\ a & x^1=z^1,x^2\neq z^2\\ b & x^1\neq z^1,x^2=z^2\\ 0 & otherwise\end{cases}$$

for some $a,b\in(0,1)$.

We can now state:

Observation 2 *Let there be given $s:X\times X\to\mathbb{R}_+$ and $\beta\leq 1$. There exists a model $\varphi_{s,\beta}$, such that $\varphi(\mathcal{A}\backslash\mathcal{CB})=0$ and for every history h_t and every $y\in Y$,*

$$S(h_t,y)=\varphi_{s,\beta}(\mathcal{A}(h_t,\{y\})).$$

This observation is verified by constructing the model

$$\varphi_{s,\beta}(\{A_{i,t,x,z}\})=\beta^{(t-i)}s(x,z). \tag{5}$$

At history $h_t=h_t(\omega)$, only the conjectures $\left\{A_{i,t,\omega_X(i),\omega_X(t)}\mid i<t\right\}$ yield predictions that are included in a singleton $\{y\}$. Hence, of all the case-based conjectures, only t conjectures will affect the prediction, corresponding to the t possible conjectures of the form $A_{i,t,\omega_X(i),\omega_X(t)}$ (with $i=0,1,...,t-1$). Moreover, this implies that the set of all relevant case-based conjectures, at h_t, $\mathcal{CB}(h_t)$ is in $\mathcal{E}$. These conjectures will be divided among the $|Y|$ possible values, each lending its weight to the outcome that occurred at the corresponding period i , $\omega_Y(i)$.

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 $(i-k+1),...,i$ are identical to observations $(t-k+1),...,t$, and so forth.

3.3 Rule-Based Reasoning

The earliest models of reasoning employing general rules date back to Greek philosophy and its study of logic, focusing on the process 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 [27], McDermott and Doyle [28], Reiter [32]), probabilistic (Nilsson [29]), and a variety of other new logics.[18]

We do not have a precise definition or rule-based reasoning. Instead, rule-based reasoning is to a large extent a collection of everything that is left over after one extracts whatever categories are of interest. We illustrate in this section some examples of reasoning that we regard as rule-based.

In many circumstances, we will think of particular conjectures as capturing rules. For example, 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

$$A = \{\omega \in \Omega \mid \omega_{X^2}(t) = 0 \quad or \quad \omega_Y(t) = 1 \quad \forall t\}. \tag{6}$$

(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 "A implies B" is then read as "A is false, or B is true, or possibly both.")

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 (6) 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. Specifically, if we consider a period t in which the level of conflict did *not* rise, the antecedent of rule A does not hold ($\omega_{X^2}(t) \neq 1$). This ensures that any value $\omega_Y(t)$ is compatible with A, and hence that the weight of the rule $\varphi(\{A\})$ will

[18] See also Gardenfors [13] and Levi [25].

not be counted in the summation $\varphi(\mathcal{A}(h_t, Y'))$ for any $Y' \subsetneq Y$. In general, 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

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

Holland's [20] genetic algorithms employ additive aggregation over rules. This method addresses a classification problem where the value of y is to be determined by the values of $x = (x^1, ..., x^m)$, based on past observations of x and y. The algorithm maintains a list of association rules, each of which predicts the value of y according to values of some of the x^j's. For instance, one rule might read "if x^2 is 1 then y is 1" and another, "if x^3 is 1 and x^7 is 0 then y is 0." In each period, each rule has a weight that depends on its success in the past, its specificity (the number of x^j variables it involves) and so forth. The algorithm chooses a prediction y that is a maximizer of the total weight of the rules that predict this y and that apply to the case at hand.

The prediction part of genetic algorithms is therefore a special case of our framework, where the conjectures are the association rules involved. However, in a genetic algorithm the set of rules does not remain constant, with rules instead being generated by a partly-random process, including crossover between "parent genes," mutations, and so forth.

We may be interested in sets of rules, and in tracking how the agent's weight of credence shifts among them. For example, suppose the agent is convinced that the price of oil will steadily increase over the next fifty hears (and has no idea what will happen after that, believing that at that point the world's reserves of oil will be exhausted). The she may attach her highest weigh to credence to those states that feature 1's in their first fifty places. She may attache her next highest weight to the collection of states that feature a single 0 in the first fifty periods, followed by a somewhat lower weight attached to states featuring two 0's, and so on. As time progresses, her prediction will be dominated by the unfalsified conjecture featuring the smallest number of 0s, and hence she will continue to predict increases.

There are many examples of rule-based reasoning that go beyond the simple cases we have just discussed. Indeed, *any* rule with a clear empirical meaning corresponds to a conjecture A, which is its extension: the set of states of the world that are consistent with the rule.

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, to foresee all possible eventualities and assign probabilities to them. The agent has a probability p over the states of the world, Ω. However, in the back of her mind she also carries with her some general rules and analogies. Assume that she employs a model φ such that

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

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

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

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

If ε is small, the non-Bayesian conjectures will play a relatively minor role is determining predictions, as long as history proceeds along a path that had a high a-priori probability p. However, suppose that the reasoner faces an eventuality, such as the September 11 attacks or the Lehman Brothers' collapse, that is surprising in the sense that the agent had assigned low or even zero probability p to this event. How will the agent make predictions? If she has assigned the event zero probability, Bayesian updating will not be well-defined. In this case, the non-Bayesian conjectures, whose total weight is bounded by ε, 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.

If the event in question had a nonzero but very small prior probability, non-Bayesian reasoning will again be relatively more important. For example, it is possible that the agent has conceived of the possibility of Lehman Brothers' collapse, but assigned a very small probability to this event. Once the event occurred, conditional probabilities are well-defined and can be used. However, non-Bayesian conjectures, which used to have a negligible effect on the reasoner's predictions, 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 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 assessment, 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."

Our framework can thus describe the reasoning of agents who are mostly Bayesian most of the time. However, they have a certain degree of self-criticism that allows them to doubt their probabilistic assessments when they encounter surprises. Indeed, as we will see in the next section, it may not be easy for the reasoner to try to avoid surprises and at the same time to remain Bayesian.

4 Dynamics of Reasoning Methods

4.1 Bayesian vs. non-Bayesian Reasoning

We first illustrate the unified model with a positive question. Under what conditions will Bayesian reasoning survive as evidence accumulates, and when will the agent turn to other modes of reasoning? Our answer is that Bayesian reasoning will wither away if the agent's prior is not sufficiently informative.

4.1.1 Assumptions

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

Assumption 1

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

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

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

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

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 $h_t, h'_t \in H_t$,

$$\frac{\varphi\left(\mathcal{B}(h_t)\right)}{\varphi\left(\mathcal{B}(h'_t)\right)} < M. \tag{7}$$

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 (Ignorance) *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 will be violated if, as often assumed in Bayesian models, the agent believes she faces successive iid draws, say, $\omega_Y(t) = 1$ in each period with probability $p > 0.5$.[19] In this case the agent believes that she knows

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

the data generating process up to the specification of a single parameter. This is not the case we are interested in. Rather, Assumption 2 is designed to capture the intuition that the agent does not know the data generating process.[20]

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

$$\frac{\varphi(\{A_{i,t,x,z}\})}{\varphi(\{A_{i',t',x',z'}\})} \leq Q(t). \tag{8}$$

However, suppose (analogously to (5)) that there exists a similarity function $s : X \times X \to \mathbb{R}_+$, a decay factor $\beta \in (0, 1]$, and a constant $c > 0$ such that, for every $i < t$ and every $x, z \in X$,

$$\varphi(\{A_{i,t,x,z}\}) = c\beta^{t-i}s(x, z). \tag{9}$$

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 (8) if $\beta < 1$, as the relevance of older vintages then declines exponentially. Fortunately, there is an obvious and easily interpretable generalization of (8) that allows us to encompass (9).

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 $t' < t$,*

$$\frac{\varphi(\{A_{i,t,x,z}\})}{\varphi(\{A_{i',t',x',z'}\})} \leq Q(t) \tag{10}$$

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

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

[20]Interestingly, violations of Assumption 2 do not necessarily mean that the conclusion of the following result is false. See the discussion in Section 4.1.3.

Condition (10) 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 (11) 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.

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

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

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 this 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? This increase of the posterior probability of the true state does not result from any change in its prior probability, but from the exclusion of falsified states. In other words, the conditional probability of the true state increases at an exponential rate because its denominator, given by 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.

The key observation is that, as t grows, and under the ignorance assumption, the weight of Bayesian conjectures that remain unrefuted by history h_t, $\varphi\left(\mathcal{B}(h_t)\right)$, becomes an exponentially small fraction of the original weight of all Bayesian conjectures, $\varphi\left(\mathcal{B}\right)$. In contrast, the number of case-based conjectures at period t is only a polynomial (in t), and there is no reason for the

weight of those that make predictions at history h_t to decrease exponentially fast in t. Unless one explicitly builds into φ an exponential decline of the weight of case-based conjectures (which is ruled out by Assumption 3), the *relative* weight placed on Bayesian conjectures 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 extolling the virtues of case-based 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 predicting the most recent outcome. Similar to kernel classification, the case-based predictions we consider here further generalize these prediction by allowing different past observations to have different weights given their similarity to the current period. While 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.

The proposition is driven by the fact that there are fewer case-based conjectures than there are Bayesian ones. More generally, when there are fewer conjectures in a given class, each of them gets a larger share of the credence weight. However, should a smaller class of conjectures have unrefuted representatives at each history, it must be the case that many of these conjectures make no predictions at many histories. Thus, classes of conjectures that are often silent may retain a higher weight than classes of conjectures that make predictions most of the time. In a sense, conjectures of the former type may be viewed as saving their ammunition and picking their fights selectively. When such a conjecture makes a prediction, it gets a relatively high weight, as if it were getting credit for the meta-knowledge, knowing when to predict and when to remain silent.

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 latter 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 refocuses the question in terms of the a priori assignment of weights: Should we not make the weight of case-based conjectures decline exponentially fast in t (violating Assumption 3), to give the Bayesian ones a fair chance, as it were? We believe that, when all Bayesian conjectures get similar weights, the answer is negative. To see why, assume first that there were only one Bayesian conjecture with a positive φ. In this case, at time t, if this conjecture is still unrefuted, the agent might 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. This, however, is not the case when Assumption 2 holds. Under this assumption, when all Bayesian conjectures get similar weights at the outset, there is no wonder 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. Hence it seems wrong to artificially increase the relative weight of these conjectures as if they were a priori selected. Decreasing the weight of the case-based conjectures at an exponential or even faster rate would be tantamount to a pre-commitment to the alternative, Bayesian approach, irrespective of how successful it will indeed be in its predictions.

4.1.3 When will Bayesianism Prevail?

Bayesian reasoning is a common and successful method of learning. This suggests that there are many learning problems in which some of the assumptions of Proposition 2 do not hold. We are therefore led to ask, under which alternative assumption will Bayesian reasoning remain useful in the long run, or even dominate other reasoning methods?

Clearly, an agent who is committed to Bayesianism (i.e., who assigns $\varphi(\mathcal{B}) = 1$ contrary to Assumption 1) will remain Bayesian. Our interest is in agents who satisfy Assumption 1 and for whom the relative weight of the Bayesian conjectures remains large or increases over time. We consider

several examples.

Example 1 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 2 A slightly less trivial example is the following. Suppose the agent believes she faces a cyclical process, but is uncertain of its period. To capture these beliefs in a simple model, let us consider only Bayesian and case-based reasoning. In addition, let $X = \{0\}$ and $Y = \{0, 1\}$, so that all periods have the same observable features, and they only differ in the binary variable the agent is trying to predict. For $k \geq 1$, let $\omega^k \in \Omega$ be defined by

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

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

Let the agent's beliefs satisfy

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

and

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

Thus, the agent splits all the weight of the Bayesian conjectures among the conjectures $\{\omega^k\}$ and leaves no weight to the other Bayesian beliefs.[21] 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

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

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

Example 3 Let us again take $X = \{0\}$ and $Y = \{0, 1\}$, and restrict attention to Bayesian and case-based reasoning. Suppose the agent believes that values of Y are independently and identically distributed across periods, with a probability of $\omega_Y(t) = 1$ that is unknown and drawn from a finite set, for simplicity taken to be the set $\left\{\frac{1}{4}, \frac{3}{4}\right\}$. The resulting process violates Assumption 2. However, for every h_t, the weight of the Bayesian conjectures consistent with h_t decreases exponentially fast in t. This suggests that a result analogous to Proposition 2 will still hold. Specifically, one may replace 2 by the more general condition, that there exists $\gamma < 1$ such that, for some polynomial $P(t)$, for every t and every h_t,

$$\varphi\left(\mathcal{B}(h_t)\right) \leq \gamma^t P(t) \tag{12}$$

and conclude that the ratio $\varphi\left(\mathcal{B}(h_t)\right) / \varphi\left(\mathcal{CB}(h_t)\right)$ still converges to 0 at each ω. (Indeed, the first part of the proof of Proposition 2 consists in showing that Assumption 2 implies this condition.) This condition holds in the case of an iid Bernoulli random variable as long as its parameter is known to be bounded away from $0, 1$. Thus, for Bayesian reasoning to survive in this set-up, the agent has to make sure that the case-based conjectures get exponentially decreasing weight, as mentioned above. Alternatively, it seems more reasonable to suggest that the agent predict an average of the realizations of the random variable, rather than exact sequences thereof.

Example 4 Considering the same set-up, $X = \{0\}$ and $Y = \{0, 1\}$, let us limit attention to the first T periods. Consider a Bayesian agent who has a uniform belief over the average

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

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

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

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

4.2 Case-Based vs. Rule-Based Reasoning

This section sketches another positive application of the model, dealing with the dynamics of case-based versus rule-based reasoning. In a sense, Example 4 above can also be viewed as comparing the two modes of reasoning. However, in that example that "rule" and its exceptions were jointly modeled by Bayesian beliefs. Here our focus is on "rules" that are given directly by multi-state conjectures.

Consider again the simplest case of $X = \{0\}$, $Y = \{0, 1\}$. Assume that y_t are iid, where $y_t = 1$ with probability p.

Define the set of rules,

$$\mathcal{RB} = \{R_{i,y} \mid i \geq 0, y \in Y\},$$

where

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

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

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

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

Thus, both $\mathcal{RB}$ and $\mathcal{CB}$ are countable. At history h_t there are precisely t case-based (unrefuted and non-tautological) conjectures, that is, $\mathcal{CB}(h_t)$ contains t conjectures, whereas the number of rule-based conjectures in $\mathcal{RB}(h_t)$ ranges between 2 and $(t+1)$.

When recent history is suggestive of a simple rule (a large number of observations of 0 or of 1), the agent adopts the rule "recent observations will continue forever." When recent history is more spotty, and no simple rule explains it, the agent assigns less weight to rule-based reasoning and resorts to case-based reasoning, which in this case means reliance on past frequencies. Since, for every k, there is a positive probability to observe a run of k 0's or k 1's, in the long run we should expect to find periods in which history suggests rules, followed by periods in which no rule seems to explain the data. Therefore, it should be expected that from time to time there will emerge a theory that is accepted by most agents, and at some point it will collapse. When it does collapse, confusion may lead agents to adopt less theoretical, more case-based methods, until the data seem to suggest a new theory, and so forth. 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 rule-based 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. Theories or rules are exciting when they succeed, but, being ambitious, they can also fail. Cases, by contrast, are no more than an amalgamation of

data, and thus they do not provide any deep insights or a sensation of "understanding." On the bright side, they can never be refuted. They are always there, waiting faithfully for the agent, who would devote more attention to them when her heroic attempts to understand the process fail.

5 Optimal Credence

We not turn to a normative question. Suppose that the agent can choose her weighting function φ. How should she make this choice? Or, what should be the criterion for choosing φ?

In this set-up, the agent has no prior probability over the state space. In fact, the question we ask is precisely which is an appropriate "prior probability,"where the latter may put some weight of credence on multi-state conjectures. In the absence of such a prior, expected utility maximization is not well-defined. Instead, one may consider the maxmin payoff (or minmax loss) or minmax regret criteria.

The agent will always be safer predicting that the outcome will be in larger rather than smaller sets $Y' \subset Y$. To make the prediction problem meaningful, suppose that at each history h_t, the agent must predict a single element of Y. The agent's decision rule is to predict a value y' that maximizes the function $\varphi(\mathcal{A}(h_t, \{y\}))$.

Now let

$$\mathcal{L}(\varphi, \omega\}$$

be the expected loss the agent encounters if she uses weighting function φ and the actual state is ω. We take this to be the discounted sum of the agent's losses over her infinite horizon, for some discount factor less than one.[22]

Given a state ω and model φ, the agent's payoff is

$$-\mathcal{L}(\varphi, \omega\}$$

[22]This is in general an expected payoff because histories may occur at which $\varphi(\mathcal{A}, \{y\})$ has multiple maximizers, in which case we would like to allow the agent to randomize. We could extend this formulation to allow the agent to make multivalued predictions, with a loss from prediction $|Y'|$ of $1/|Y'|$ if the realized outcome lies in $|Y'|$ and a loss of 1 otherwise, without affecting the result. We could substitute a limit-of-the-means payoff criterion for the discounted expected loss, at the cost of some technical modifications in the arguments, to account for the fact that changes in behavior over any finite number of periods are irrelevant for a limit-of-the-means calculation.

and her regret is

$$\mathcal{L}(\varphi, \omega) - \inf_{\varphi'} \mathcal{L}(\varphi', \omega).$$

Now suppose the agent chooses φ to either to minimize her maximum loss

$$\sup_{\omega \in \Omega} \mathcal{L}(\varphi, \omega)$$

or to minimize her maximal regret:

$$\sup_{\omega \in \Omega} \{\mathcal{L}(\varphi, \omega) - \inf_{\varphi'} \mathcal{L}(\varphi', \omega)\}.$$

This turns out to be an unhelpful criterion for choosing φ:

Proposition 3 *For any φ,*

$$\sup_{\omega \in \Omega} \mathcal{L}(\varphi, \omega) = \sup_{\omega \in \Omega} \{\mathcal{L}(\varphi, \omega) - \inf_{\varphi'} \mathcal{L}(\varphi', \omega)\} \geq \frac{|Y| - 1}{|Y|}.$$

There exists a function φ that achieves the lower bound $\frac{|Y|-1}{|Y|}$ on the minmax loss or minmax regret, and the only such function is one that after every history h_t sets $\varphi(\mathcal{A}(h_t, \{y\})) = \varphi(\mathcal{A}(h_t, \{y'\}))$ for all $y, y' \in Y$.

The implication of this result is that the agent minimizes her maximum loss, or minimizes her maximum regret, by adopting a model that is absolutely useless in prediction.

This result should not be too surprising: in the absence of additional structure, no learning is possible. For learning to be effective, one needs to know something about the problem, for instance, that the data generating process is one of a given set of distributions. But if nothing is known, and nature, as it were, is allowed to be malevolent and choose a state that is worse for the agent's prediction strategy, the agent will be better off by random predictions than by reasoned ones.

6 Concluding Remarks

6.1 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, that is, the argument of φ, to be sets of conjectures, with the understanding that within each set a most successful conjecture is selected for prediction, and that the degree of success of the set is judged by the accuracy of this most successful conjecture. The following example illustrates.

Suppose that the agent is faced with a sequence of datasets. In each dataset there are many consecutive observations, indicating whether a comet has appeared (1) or not (0). Different datasets refer to potentially different comets.

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

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

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

6.2 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.

6.3 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 [41] definition of induction.[23] 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.

6.4 Decision Theory

The present paper deals with prediction. In order to explore its implications 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

[23] See Solomonoff [39], who suggested to couple this preference for simplicty 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.

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.

7 Appendix: Proofs

7.1 Proof of Proposition 1

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

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

for every conjecture of the form $\{[h_t, Y'] \cup (h_t)^c\}$, and set $\varphi(\mathcal{F}) = 0$ where $\mathcal{F}$ is the set of all conjectures that are not of this form, and $c_{h_t} > 0$ is to be determined. Observe that the conjecture $[h_t, Y'] \cup (h_t)^c$ is unrefuted 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

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

■

7.2 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

$$\varphi(\mathcal{B}(h_t)) \leq \frac{P(t)}{d^t}. \tag{13}$$

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

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

Then,

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

Choose T large enough so that

$$\sum_{\tau=1}^{T} S_\tau > \frac{\varphi\left(\mathcal{CB}\right)}{2}. \tag{14a}$$

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

Consider a conjecture $A_{(t-1),t,x,z} \in \mathcal{CB}$ and assume that $\varphi\left(\left\{A_{(t-1),t,x,z}\right\}\right) < \xi$. By (10) (of Assumption 3) we have that, for all $t' < t, x', z'$

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

By (11) (of that Assumption), we know that for all $i < t' < t$, and all x', z',

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

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

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

and, using (14a) we obtain

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

Define

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

and observe that it is a polynomial in t.

Thus, we have

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

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

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

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

$$\varphi\left(\mathcal{CB}(h_t)\right) \geq \varphi\left(\left\{A_{(t-1),t,x,z}\right\}\right) \geq \frac{\varphi\left(\mathcal{CB}\right)}{R(t)}. \tag{15}$$

Combining (13) and (15) we obtain

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

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

■

7.3 Proof of Proposition 3

Fix a state ω, and then consider a function φ with $\varphi(\{\omega\}) = 1$ and $\varphi(A) = 0$ for all other conjectures. Then $\mathcal{L}(\varphi, \omega) = 0$. This in turn ensures that $\inf_{\varphi'} \mathcal{L}(\varphi', \omega) = 0$ for all ω, and hence that the minmax loss and minmax regret criteria are identical.

We now show that

$$\max_{\varphi}\left[\sup_{\omega\in\Omega}\mathcal{L}(\varphi,\omega)\right] = \frac{|Y|-1}{|Y|}$$

Consider a two-person zero-sum game between the agent, choosing φ, and nature, choosing ω, where the agent's payoff is $-\mathcal{L}(\varphi, \omega)$. Fix φ and define $\omega\left(\varphi\right)$ inductively as follows: at each $h_t(\omega)$, choose the y value that has the lowest weight $\varphi(\mathcal{A}(h_t, \{y\}))$. The probability that the agent, predicting a y in $\arg\max_y \varphi(\mathcal{A}(h_t, \{y\}))$ will be correct is bounded by $\frac{|Y|-1}{|Y|}$, and therefore so is the long-run average payoff, $\mathcal{L}(\varphi, \omega)$. Since the payoff is discounted, one may invoke the maxmin theorem to conclude that $\frac{|Y|-1}{|Y|}$ is an upper bound on $\sup_{\omega\in\Omega} \mathcal{L}(\varphi, \omega)$ for every φ.

Clearly, a random φ obtains this bound. For any other φ', such that $\arg\max_y \varphi(\mathcal{A}(h_t, \{y\}))$ is a proper subset of Y for some h_t, $\omega\left(\varphi'\right)$ achieves a

strictly lower payoff (than $\frac{|Y|-1}{|Y|}$) at h_t, and therefore also for the discounted average. ■

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

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

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

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

[31] 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.

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

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

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

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

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

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

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

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

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

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