A Unified Model of Induction[1]

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

October 18, 2009

Abstract

We suggest a model of inductive inference that includes, as special cases, Bayesian reasoning, case-based reasoning, and rule-based reasoning. Our unified approach allows us to examine how the various modes of reasoning can be combined and how the weight of credence can shift among them. We establish conditions under which a reasoner 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 reasoning relative to that put on case-based and rule-based reasoning. We show that case-based reasoning can help mitigate problems of overfitting the data. Finally, we show that random data may make certain theories seem plausible, and, as a result, increase the weight of rule-based vs. case-based reasoning.

[1]We thank Eddie Dekel, Gabi Gayer, Offer Lieberman, and George Mailath for comments and suggestions. Financial support from the National Science Foundation (grants SES-0549946 and SES-0850263) is gratefully acknowledged.

[2]Tel Aviv University, HEC, Paris, and Cowles Foundation, Yale University.

[3]Department of Economics, Yale University

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

Contents

A Unified Model of Induction

Itzhak Gilboa, Larry Samuelson, David Schmeidler

October 18, 2009

1 Introduction

Induction is the process of learning from the past about the future, or more generally, of learning from given data what might be predicted about yet unobserved cases. Understanding how people perform induction is a key component in studying individual and social behavior, underlying much of the study of psychology, sociology and economics. Studying *optimal* ways of performing induction is at the heart of the philosophy of science, statistics, artificial intelligence and machine learning.

Three basic modes of reasoning, with roots nearly as old as recorded history, have been used for inductive learning: case-based (analogical), rule-based (logical), and Bayesian (probabilistic). Analogical reasoning is perhaps the most primitive. Ancient texts are replete with analogies used for inductive inference as well as rhetorical purposes.[1] The generalization of particular cases into general rules appears to demand a higher analytical level. Nonetheless, such generalizations appear in ancient sources close on the heals of analogical reasoning.[2] Bayesian reasoning is perhaps the most recently formulated and least natural of the three models—there is ample evidence that highly-educated people do not excel in reasoning about probabilities.[3] Yet, the main ingredients of Bayesian thinking can be traced, in informal guises, back to early history, including (i) the distinction between degrees of plausibility, (ii) the exclusion of scenarios that are incompatible with observed evidence, and (iii) the counterfactual reasoning about what evidence one could have seen but didn't.

Analogical reasoning was explicitly discussed by Hume [24], and received attention in the twentieth century in the guise of case-based reasoning (Riesbeck and Schank [35], Schank [37]), leading to the formal models of Gilboa and Schmeidler ([19, 20, 21]). The earliest models of rule-based reasoning date back to Greek philosophy and its study of logic. The rise of analytical philosophy, the philosophy of mathematics, and, artificial intelligence

[1] See, for instance, Browning [5].

[2] For example, the book of Proverbs in the *Bible*.

[3] See Tversky and Kahneman [42].

greatly extended the scope of rule-based reasoning, including the introduction of non-monotonic logics (McCarthy [28], McDermott and Doyle [29], Reiter [34]), probabilistic logics (Nilsson [30]), and a variety of other new logics.[4] Bayesian reasoning appeared explicitly in the writings of Bayes [3],[5] and has received much attention in the twentieth century. Beginning with the work of de Finetti and his followers, it has given rise to the Bayesian approach to statistics. It has grown from the early work of Ramsey [33], de Finetti [10], and Savage [36] to become the dominant approach in economic theory and in game theory. Within the philosophy of science, proponents of the Bayesian approach include Carnap [7], Lindley [27], and Jeffrey [25]. Finally, the Bayesian approach has achieved great success in computer science and artificial intelligence, such as in the context of Bayesian networks (Pearl [31]).

This paper presents a model that unifies these three modes of induction (and potentially others), allowing us to view them as special cases of a general learning model. Our approach is illustrated in Section 2, while the formal setting and model of induction are developed in Sections 3 and 4. This unified model has several uses. First, as we see in Section 4, putting the three common models of induction on a common footing gives us a deeper understanding of each, as well as of the relationships between them. Second (Section 5), a unified model allows us to study the dynamics of reasoning: when do people tend to use Bayesian reasoning more, and when would they tend to use rules, or cases? What affects these choices, and what patterns can we observe in terms of switching among modes of reasoning? Third, we can ask when it makes sense to be a Bayesian (Section 5.3). While it is standard throughout economic theory to assume that people are Bayesian, it is well understood that this assumption is not always sensible. Our unified treatment can help identify scope of applicability of the Bayesian model.

2 Toward a Unified Model: An Example

All three types of reasoning can typically be applied to a given induction problem, and often give rise to similar conclusions. For example, a case-based reasoner might say, "The sky is filled with black clouds. I can recall many such days on which it has rained, and not so many when it has not.

[4]See also Levi [26] and Gardenfors ([14].

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

Therefore, I predict that it will rain today." A rule-based reasoner might address the same problem by saying, "I have observed that, when the sky is cloudy, it tends to rain. Given the clouds I see, I predict rain." Finally, the Bayesian approach to the same problem might be, "I had a prior probability distribution about the generation of (independent and identically distributed, or more generally exchangeable) joint observations of clouds and rain. Having updated this prior on the basis of my past observations, I attach a high conditional probability to rain, given the current cloudy sky." How can we tell these modes of reasoning apart, and make the differences between them precise?

Our model begins with the standard decision-theoretic notion of a set of states of the world, describing all scenarios that might unfold. The basic ingredient of our model is then the notion of a "hypothesis." A hypothesis will be formally modeled as an event, i.e., a subset of states of the world. We assume that the reasoner assigns a certain non-negative "weight of credence" to each hypothesis of which she can conceive. Others are assigned a zero weight in our model.

Given a history of observations, some hypotheses will be refuted. We assume that the reasoner excludes all such refuted hypotheses from her analysis. We also assume the reasoner ignores hypotheses that are consistent with *any* possible current outcome and hence provide no clue as to what the reasoner might predict. Each of the remaining hypotheses generates some predictions, namely, the (possibly many, but not all) observations that may follow the observed history according to the hypothesis in question. A hypothesis is assumed to provide its weight of credence as support for all the predictions consistent with it. These weights of credence are aggregated by simple summation. One prediction is then more likely than another if the total weight of credence of the former (summing over all hypotheses that support it) is higher than that of the latter.

This benchmark model has several variants. It can be used to generate qualitative predictions, simply ranking the possible predictions, or quantitative ones, taking the form of a probability distribution over the possible observations. The linear aggregation can allow many hypotheses to affect the prediction, or in a limiting case can base the prediction on a single, "most plausible" hypothesis.

The flexibility of the model stems from the freedom allowed by the concept of a hypothesis and by the assignments of the weights of credence to hypotheses. On the one extreme, a hypothesis can be a general rule or sci-

entific theory. In this case, the hypothesis itself may capture much of the agent's reasoning about the world, and aggregation of hypotheses may be relatively important. At the other extreme, a hypothesis may be a single state of the world. Such hypotheses clearly cannot be considered scientific theories, but they are used to capture the agent's reasoning by aggregation over the weights of hypotheses that concur in terms of their predictions. If *all* hypotheses are of this type, our model reduces to a Bayesian model, in which the weight of credence of a hypothesis is simply the probability of the corresponding state of the world.

Conditional statements can be captured in this model by mapping a statement of the type "if p then q" to the collection of all states (i.e., the hypothesis) in which either $\{p$ and $q\}$ or $\{\neg p\}$ (i.e., not-p). If $\neg p$, then the hypothesis holds whether q or $\neg q$, and is excluded from consideration by the convention that only hypotheses with nontrivial implications for prediction are taken into account. If p, then the hypothesis is among those that are both compatible with past observations and relevant for current predictions, and it is therefore used for prediction.

To illustrate the model, suppose that in each period $0 \leq t \leq T-1$, we observe whether the sky was cloudy (i.e., whether $x_\tau \in \{0,1\}$) and whether it rained (whether $y_\tau \in \{0,1\}$) in each of the preceding periods, as well as whether it is currently cloudy (x_t), and we are then asked to predict the precipitation y_t. The set of all states of the world is therefore all sequences of T pairs of the type $(x_t, y_t) \in \{0,1\}^2$.

A Bayesian reasoner in this setting has a probability for each state of the world. In particular, for each state ω she will have a hypothesis $\{\omega\}$, and she assigns to this hypothesis a weight of credence that is equal to the probability of the state. Assigning zero weight to all other hypotheses (that is, to all events that have more than one state), the aggregation over hypotheses is equivalent to the addition of probabilities over states. For example, if asked at time t whether y_t is more likely to be 0 and 1, a standard Bayesian reasoner excludes those states inconsistent with the observed history $((x_0, y_0), ..., (x_{t-1}, y_{t-1}))$, normalizes the probability on the remaining states to get a total probability of 1, and then asks whether the sum of the probabilities attached to states in which $y_t = 0$ is larger or smaller than the sum for states in which $y_t = 1$. The reasoner in our model would do essentially the same, dispensing with the normalization step: she will exclude those single-state hypotheses inconsistent with the data, and compare the sum of the weights of credence attached to hypotheses in which $y_t = 0$ to the

corresponding sum for hypotheses in which $y_t = 1$.

Consider now a simple case-based reasoner, who expects to observe the outcome that has occurred in many similar cases in the past. Faced with x_t, the standard model of a case-based reasoner asks her to go back and find all previous cases with similar x values, while ignoring all other, dissimilar cases. She would then find the value $y \in \{0, 1\}$ that was more common among the collection of similar cases, and announce it as her prediction.

In our model, such a reasoner can be captured as considering only hypotheses of the type "case i and case t have x values equal to x_i and x_t, respectively, and they have the same y value." It is as if we thought of the case-based model as a set of conditional statements, each of the type, "If we observe x_i in period i and x_t in period t, then we will observe the same y values in both periods"—where we have such a conditional hypotheses for each pair of values x_i and x_t (and each pair of periods i and t). The weight on each such hypothesis will be the similarity between x_i and x_t.

In our example, where x can only take two values, the notion of "similarity" is not particularly interesting: we must either let a value x be similar (or dissimilar) to *all* possible values x', or we are left with a similarity relation that is the identity. Yet, this simple example conveys the general idea. The reasoner will have observed the actual x_i and x_t upon arriving at period t, and will exclude the many hypotheses that do not specify these values. As a result, under the identity similarity function appropriate for our clouds-and-rain scenario, the period-t prediction will be determined by past cases for which $x_i = x_t$. For each such past case, the corresponding hypothesis predicts the value $y_t = y_i$. In other words, the reasoner in our model considers the observations $(y_i)_{\{i|x_i=x_t\}}$, and evaluates the plausibility of a prediction $y_t \in \{0, 1\}$ according to the sum of similarity values between the present case and past such cases that resulted in the occurrence of the same y_t.

Observe that the hypotheses that the case-based reasoner entertains are much larger than those of the Bayesian: whereas the latter consider hypotheses that dictate the values of (x_t, y_t) for all t, and contain but one state of the world each, the former reasons in terms of hypotheses that restrict states in two periods only, and contain 4^{T-2} states each.

Finally, a rule-based reasoner might be modeled as associating weight to hypotheses of the form "whenever $x_t = 1$, then $y_t = 1$," corresponding to the rule "it always rains when the sky is cloudy" in our example. Such a hypothesis is more universal and hence more restrictive than the case-based hypotheses, as it potentially makes a prediction about (and can be refuted

by) the observations in each and every period. At the same time, it is not as specific as a Bayesian hypothesis, as it remains silent about the prediction in certain periods (when $x_t = 0$).

Another rule-based hypothesis might be "for every t, $y_t = 1$," reflecting the claim "it always rains." This hypothesis definitely provides a prediction at every period. In our model, if this hypothesis is assigned a high initial weight, it will have a significant effect on the prediction as long as it is not proven wrong. Once it is refuted, it will disappear from future considerations, leaving the stage to other hypotheses. In particular, the reasoner may start with positive weights both on the unconditional rule, "for every t, $y_t = 1$", and on the conditional one, "whenever $x_t = 1$, then $y_t = 1$". If at a given period t she finds $x_t = 0$ and $y_t = 0$, the former hypothesis will be eliminated while the latter will continue to take part in generating predictions.

The only process of learning in our model consists of excluding hypotheses that have been refuted by the data. For a Bayesian reasoner, this is equivalent to Bayesian updating; for a case-based reasoner, this process selects the relevant past cases; and for a rule-based reasoner, excluding refuted theories resembles the fundamental process of scientific progress. However, this basic model of induction can do more than mimic these known forms of learning. It can also shift weight among the various modes of reasoning. For example, when a rule is excluded from further predictions, the entire weight of rule-based hypotheses also decreases. We may therefore find that the collapse of a particular theory implies not only that other theories will come to the fore, but also that the entire rule-based reasoning method loses weight as compared to, say, a case-based one. Similarly, if we compare the total weight assigned to Bayesian hypotheses (i.e., states of the world) and that assigned to case-based hypotheses, under fairly general conditions the former decreases at an exponential rate, while the latter decreases at only a polynomial rate. As a result, a reasoner who starts off with a potentially small but positive weight assigned to the case-based mode of reasoning will converge to reasoning mostly by analogies, and to putting relatively no weight on Bayesian reasoning. We will argue that this result is rather robust, unless the reasoner has some a-priori knowledge about the nature of the process she is facing.

3 The Induction Problem

3.1 The Environment

Each period $t \in \{0, 1, 2, \ldots, T-1\} \equiv \mathbb{T}$ features a *characteristic* $x_t \in X$ and an *outcome* $y_t \in Y$. The sets X and Y are finite and non-empty.[6] The set of all *states of the world* is

$$\Omega = \{\omega : \mathbb{T} \to X \times Y\}.$$

We let $\omega(t) = (\omega_x(t), \omega_y(t))$ be the element of $X \times Y$ appearing in period t given state ω, and then let

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

denote the history of characteristics and outcomes in periods 0 through $t-1$ given state ω. The set of all such histories is

$$H_t = \cup_\Omega h_t(\omega) = (X \times Y)^t.$$

We are also interested in histories of t periods, coupled with period-t characteristic $\omega_x(t)$. Let

$$h_t^*(\omega) = (\omega(0), \ldots, \omega(t-1), \omega_x(t))$$

denote such a history, and let the set of all such histories be given by

$$H_t^* = \cup_\Omega h_t^*(\omega) = (X \times Y)^t \times X.$$

For $h_t^* \in H_t^*$ and $y \in Y$, let (h_t^*, y) be the element of H_{t+1} obtained from concatenating y to the last element of h_t^*.

3.2 Prediction

In each period $t \in \mathbb{T}$, the reasoner observes a history h_t^* and must make a prediction about the period-t outcome, $\omega_y(t) \in Y$. A *qualitative* prediction consists of a ranking of the values in Y given h_t^*, that is, a binary relation $\succsim_{h_t^*} \subset Y \times Y$. A *quantitative* prediction is a probability over the set Y, i.e., $p_{h_t^*} \in \Delta(Y)$.

[6] No conceptual problems arise in extending the analysis to infinite sets X, Y or $\mathbb{T}$, but we avoid a collection of technical complications by working with finite sets.

The reasoner may make this prediction with the help of hypotheses. A *hypothesis* is an event $A \subset \Omega$. A hypothesis can represent a theory, an association rule, an analogy, or in general any reasoning aid one may employ in predicting y_t. Indeed, any such reasoning tool can be described extensively, by the set of states that are compatible with it. This set is the event used to describe the hypothesis in this model. The set $\mathcal{A}$ of all hypotheses is given by

$$\mathcal{A} = \{0,1\}^{\Omega}.$$

3.3 Compatibility and Relevance

Two basic distinctions between hypotheses are useful. First, we distinguish the hypotheses that are compatible with the data the reasoner has observed from those that aren't. Second, among the unrefuted hypotheses, we distinguish those that are useful from those that are useless for the prediction problem at hand.

For a history $h_t \in H_t$ or $h_t^* \in H_t^*$, we define the events $[h_t]$ and $[h_t^*] \subset \Omega$ respectively as all the states that are compatible the corresponding history, that is,

$$\begin{aligned} & [h_t] = \{\omega \in \Omega \,|(\omega(0), \ldots, \omega(t-1)) = h_t\} \\ \text{and} \quad & [h_t^*] = \{\omega \in \Omega \,|(\omega(0), \ldots, \omega(t-1), \omega_x(t)) = h_t^*\}. \end{aligned}$$

We say that a hypothesis A is *compatible with* history h_t or h_t^*, denoted $A \in C(h_t)$ or $A \in C(h_t^*)$, if

$$A \cap [h_t] \neq \varnothing \quad \text{or} \quad A \cap [h_t^*] \neq \varnothing,$$

i.e., if there is a continuation of the history that yields a state in A.

We say that a hypothesis A is *relevant at* history h_t^*, written $A \in R(h_t^*)$, if $A \in C(h_t^*)$, and there exists at least one $y \in Y$ such that $A \notin C((h_t^*, y))$. Differently put, $A \in R(h_t^*)$ means that hypothesis A has not been falsified by the data available at h_t^*, but it will not vacuously hold in the $(t+1)$–st observation: it can be refuted by at least one possible observation.

An obvious requirement on a reasoning process, embodied in all of the standard models of induction, is that incompatible processes play no role in prediction. We will incorporate this requirement into our general model. We will also typically preclude unrefuted but irrelevant hypotheses from playing a role, though this restriction is less important. Specifically, the qualitative

ranking of predictions, which is the focus on this paper, will provide the same results with and without the irrelevant hypotheses.

3.4 Models

Compatibility and relevance are objective criteria, which do not depend on the reasoner's intuition, prejudices, personal tastes or biases. Unfortunately, these objective criteria alone provide no clue as to how to make predictions. The set of hypotheses is *conditionally symmetric*, in the sense that, for every history h_t^*, even after the hypotheses falsified by h_t^* have been eliminated, it is still the case that for every hypothesis that might lead one to predict y, there is an alternative hypothesis that will lead to prediction y', for any other $y' \in Y$. For example, for every hypothesis that is compatible with a sequence of observations 000...0 and leads to the prediction 0 there is another hypothesis that predicts 1 following the same sequence 000...0.[7] A reasoner can thus reason effectively only if she has an initial notion that some hypotheses are likely to be more helpful than others.

We refer to the reasoner's initial bias concerning hypotheses as the reasoner's model. Formally, a *model* is a function $\phi : \mathcal{A} \times H^* \to \mathbb{R}_+$, where $H^* = \cup_t H_t^*$ and $\phi_{h_t^*}(A) \equiv \phi(A, h_t^*)$ is interpreted as the weight attached to hypothesis A, given history h_t^*. Intuitively, a higher weight $\phi_{h_t^*}(A)$ is interpreted as suggesting that the reasoner finds hypothesis A more relevant for prediction, where $\phi_{h_t^*}(A) = 0$ might mean that the reasoner has never conceived of hypothesis A, or thought about it and found that it carries no weight.

Typically, we should expect ϕ to vanish for most hypotheses. The cardinality of the set of hypotheses $\mathcal{A}_t$ is doubly-exponential in the number of periods T:

$$|\mathcal{A}| = 2^{\left((|X|\times|Y|)^T\right)}.$$

Hence, even for moderate values of T, there are more hypotheses than one might think of.[8] The distinction between the hypotheses one has considered and those that one has not thus introduces two natural levels of weights

[7] This has been famously pointed out by Hume's [24] problem of induction. Goodman [22] raised the additional complication of the dependence of induction on language. See Gilboa [16] for a discussion.

[8] For $T = 9$, with $|X| = 1$ and $|Y| = 2$, this number of hypotheses (2^{2^9}) exceeds the estimate of the number of atoms in the observable universe (10^{80}).

one may assign to different hypotheses. We would like to go beyond this and allow the reasoner to consider some hypotheses as more useful or a priori more reasonable than others. The mental process by which the reasoner assigns positive weights $\phi_{h_t^*}(\cdot)$ to some hypotheses will depend on the context, with $\phi_{h_t^*}(\cdot)$ perhaps reflecting probability of states, similarity of cases, or simplicity of theories.

4 A General Model of Induction

4.1 Qualitative Prediction

Given a model ϕ and history h_t^*, let us extend $\phi_{h_t^*}$ to sets of hypotheses by defining, for the set $\hat{\mathcal{A}} \subset \mathcal{A}$,

$$\phi_{h_t^*}(\hat{\mathcal{A}}) = \sum_{A \in \hat{\mathcal{A}}} \phi_{h_t^*}(A). \tag{1}$$

We then assume that the reasoner's qualitative prediction ranks $y \in Y$ according to the function

$$\Phi_{h_t^*}(y) = \phi_{h_t^*}(R(h_t^*) \cap C((h_t^*, y))). \tag{2}$$

That is, faced with history h_t^*, and asking herself which of two predictions, $y, y' \in Y$ is more likely to occur, the reasoner adds up the weights of all hypotheses that are compatible with y as a continuation of h_t^* and relevant, does the same for y', and ranks the predictions y and y' in accordance with the resulting sums (cf. (2)). We thus build into the reasoner's prediction rule the fact that refuted hypotheses are excluded from further reasoning. This seems to be the most basic idea of empiricism,[9] and is satisfied in all of our examples of inductive inference. We have assumed that predictions are generated by the *sum* of the weights $\phi_{h_t^*}(A)$ for the compatible and relevant hypotheses A. Section 4.3 shows that this sacrifices no generality.

4.2 Quantitative Prediction

We might equivalently assume that the reasoner makes qualitative predictions according to

$$\phi_{h_t^*}(C((h_t^*, y))). \tag{3}$$

[9] See Carnap [6] and Popper [32].

The expression $\Phi_{h_1^*}$ given by (2) differs from the potential alternative given by (3) in that the former excludes hypotheses that are not relevant at h_t^* because they are compatible with any observation y at h_t^*. Such a hypothesis A may either add the same value $\phi_{h_t^*}(A)$ to all possible predictions y, or not add this value to all of them. Clearly, this is of no import when qualitative predictions are concerned.

The reasoner might instead make quantitative predictions. One definition of a probability distribution over Y given h_t^* is

$$\Pr(y|h_t^*) = \frac{\Phi_{h_t^*}(y)}{\sum_{y' \in Y} \Phi_{h_t^*}(y')}, \tag{4}$$

assuming that the denominator is positive. Clearly, using (3) instead of $\Phi_{h_t^*}$ in this formula will result in different probability values. A disadvantage of using $\Phi_{h_t^*}$ in (4) is that we may find the denominator vanishing too often. By contrast, using (3) will have the disadvantage that irrelevant hypotheses will be biasing the probability toward the uniform distribution over Y.

4.3 Constant Models

The content of our model of induction comes from the reasoner's model ϕ. By allowing the function ϕ to depend arbitrarily on h_t^*, we can sever all links between the different periods of the model, rendering it doubtful whether this could still be called a model of inductive inference. We will be therefore be especially interested in constant models:

Definition 1 *The model ϕ is* constant *if, for every t and $\hat{t}$ (possibly equal to t) and every pair of histories h_t^* and $\hat{h}_{\hat{t}}^*$, we have*

$$\phi_{h_t^*} = \phi_{\hat{h}_{\hat{t}}^*}.$$

A constant model thus makes no use of any additional information at time t other than identifying refuted hypotheses. We write constant models as simply $\phi(A)$ instead of $\phi_{h_t^*}(A)$, dropping the irrelevant history subscript.

At the same time that we are considering restricting the model to constant functions ϕ, one might wonder whether we shouldn't generalize the model beyond the linear aggregation allowed by (2). Suppose that the observable data consist of the rankings $\succsim_{h_t}$ for every possible h_t^*. Hence, we can observe the reasoner's complete ranking of outcomes, in every circumstance

in which the reasoner is called upon to offer such a ranking, but cannot observe additional data (such as indications of which hypotheses the reasoner is using to generate these predictions). Then we can show that there is no loss of generality in restricting attention to *both* constant functions ϕ and linear aggregation as in (2):

Proposition 1 *Let* $\Pr(y|h_t^*)$ *be a quantitative prediction. Then there exists a constant model generating* $\Pr(y|h_t^*)$.

It is immediate that the same holds for qualitative prediction.

The straightforward proof of this proposition is given in Section 7.1. It shows that, in fact, the expressive power of the Bayesian hypotheses suffices for the generation of any quantitative prediction. Indeed, many other sets of hypotheses that contain sufficiently many hypotheses (of the order of magnitude of the set of histories) may serve the same purpose. The implication of this result is that if we only observe a prediction for each possible history, $(\Pr(y|h_t^*))_{h_t^*}$, we cannot uniquely identify the function ϕ or the reasoning mode that the reasoner employs. Post-hoc, given any set of predictions, $(\Pr(y|h_t^*))_{h_t^*}$, a Bayesian function ϕ as well as many non-Bayesian functions are compatible with the predictions.

This result ensures that the linear aggregation of constant ϕ functions is sufficiently general for our purposes. At the same time, it raises doubts about the refutability of our model. If any set of predictions $(\Pr(y|h_t^*))_{h_t^*}$ can result from our model in multiple ways, what are the observable implications of the model? And if the Bayesian hypotheses suffice, why bother with additional ones?

Several points need to be made here. First, the constant, linear model is but a framework within which additional assumptions can and should be imposed. The model can be augmented by specific assumptions about the mode of reasoning to examine particular questions, as we illustrate in the next section. Second, the predictions $(\Pr(y|h_t^*))_{h_t^*}$ are not necessarily the only observable data. In everyday life as well as in scientific inquiry, the reasoning behind these predictions is also often observable: people explain how they arrive at predictions and try to justify them by reasoning techniques of the types discussed above. It is even possible that some such lines of reasoning will not be compatible with the additive model, even though the bottom line predictions will be.

Third, a note on complexity is in order. The number of hypotheses is hyper-exponential in T. Even the number of Bayesian hypotheses is exponential in T. Therefore, it does not appear reasonable that reasoners explicitly refer to all hypotheses in these sets. Rather, it makes more sense to assume that reasoners assign weight to hypotheses using algorithms that are more concise, such as dividing the weight equally among hypotheses within a certain set, or assigning positive weights only to small subsets of hypotheses. These computability considerations will render certain functions ϕ more plausible than others, and will make our general framework and its predictions more restrictive.

While we focus on constant models in the remainder of this paper, it is useful to ask, why might one be tempted to use a non-constant model? Non-constant models may be particularly well suited for examining cases in which a reasoner cannot possibly be aware of all hypotheses, but may become aware of some when the data at some point "suggest" them. In everyday life people may "discover" regularities, and in scientific research a scientist may have a sudden revelation of a new theory (cf. Aragones, Gilboa, Postlewaite and Schmeidler [2]). In one sense, these revelations may be considered to be boundedly rational: if a new hypothesis comes to mind only after observing history h_t^*, the reasoner could, in principle, imagine at period 0 what she would think were she to be confronted with this history. However, this type of bounded rationality cannot be lightly dismissed because, as mentioned above, the number of hypotheses grows at a rate that is doubly exponential in the horizon T, placing extraordinary demands on "perfectly rational" reasoners.

When the reasoner's model is constant, it will often be convenient to assume also that the weight function is normalized so that

$$\phi\left(\mathcal{A}\right) = 1.$$

Observe that this condition is imposed on the a priori weights. Once a history h_t^* has been observed, the weight of the hypotheses that are consistent with history, $\phi\left(C\left(h_t^*\right)\right)$, will typically drop below 1. We could renormalize these weight so that $\phi\left(C\left(h_t^*\right)\right) = 1$ (at each history h_t^*), as in customary in the case of Bayesian updating, but this renormalization is unnecessary and sometimes cumbersome.

Normalizing the weights $\phi(A)$ makes them look especially like subjective probabilities. We can think of Nature as initially determining a single hypothesis $A \in \mathcal{A}$ and then generating observations accordingly, with $\phi(A)$

being the reasoner's subjective probability that Nature's choice was hypothesis A. Notice, however, that such a choice does not preclude the possibility that, even after observing arbitrarily long histories h_t, the reasoner will not know which hypothesis A, out of the various hypotheses consistent with h_t, was chosen by Nature in the first stage. Hence this interpretation of ϕ as subjective probability is possible, but it is a subjective probability about facts that may never be observed in our model.[10]

In the remainder of this section we assume that predictions are given by (2), for a constant model ϕ, and show how the three types of reasoning are special cases of our model.

4.4 Bayesian Reasoning

Assume that

$$\phi(A) = 0 \qquad \text{if} \qquad |A| > 1.$$

Hence, ϕ vanishes outside of the set of *Bayesian hypotheses*, given by

$$\mathcal{B} = \{\{\omega\} \,|\omega \in \Omega\} \subset \mathcal{A}.$$

In this case the reasoner puts all the weight of credence on specific states of the world, leaving nothing unspecified. Here, $\phi(\{\omega\})$ may be viewed as the probability of state ω. After observing history h_t^*, the reasoner ignores all states that are incompatible with it. This is equivalent to Bayesian updating. More precisely, given such a function ϕ, (4) defines a probability over Y, which is updated according to Bayes rule, whenever well-defined.

4.5 Case-Based Reasoning

Assume that, for every $i < t \leq T-1$, $x, z \in X$, we have a hypothesis

$$A_{it,x,z} = \{\omega \in \Omega \,|\omega_x(i) = x, \omega_x(t) = z, \omega_y(i) = \omega_y(t)\}.$$

We can interpret this hypothesis as indicating that, if the input data in period i are given by x and in period t are given by z, then periods i and

[10]To pursue this line further, one may consider a model in which there are many repetitions of the model we discuss here, which the reasoner confronts sequentially. If Nature commits to one hypothesis, A, that should apply in all models, and is then free to choose observations in each model separately (provided that these choices are all consistent with A), the interpretation of $\phi(A)$ as the reasoner's subjective probability about Nature's first stage choice is coherent, and this choice may also be asymptotically verifiable.

t will produce the same outcome (y value). Let the set of all hypotheses of this type be denoted

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

Suppose that the reasoner places all the weight on such "case-based" hypothesis, i.e., ϕ vanishes outside $\mathcal{CB}$. The weight placed on hypothesis $A_{it,x,z}$ is given by

$$\phi(A_{it,x,z}) = s((i,x),(t,z)),$$

where s is a *similarity function* identifying how closely related are a case in which one observes x in period i and a case in which observes z in period t. Assuming that the observable variables $x, z \in X$ contain all relevant information, it is natural to assume that $s((i,x),(t,z))$ depends only on x, z. However, if the time period itself is not part of the description of the prediction problem $x \in X$, it makes sense to allow a similarity function that may decrease with the time that elapsed between the two cases. Thus, we assume

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

for $\beta \leq 1$ and a similarity function $s : X \times X \to \mathbb{R}_+$. We retain the possibility here that $\beta = 1$, in which case the time that has elapsed between periods i and t plays no role, while setting $\beta < 1$ allows us to encompass decaying memory and/or decreasing relevance of the distant past.[11]

Given history h_t^*, the aggregated prediction according to ϕ will assign to a value y the sum, over all past instances where y has been observed, of the similarity of the x values observed in these instances to the currently observed $\omega_x(t)$.

This is the most basic form of case-based prediction,[12] and it coincides with kernel classification.[13] In general, we could define similarity relations based not only on single observations but also on sequences, or on other

[11] Our results would continue to hold if we imposed only the more general condition that $s((i,x),(t,z)) = s((i+k,x),(t+k,z))$. The key implication is that the sum (from period 0 to $t-1$) of the similarities to period t not vanish exponentially as t grows. In general, one may also incorporate the time index into the description of $x \in X$, and let the similarity function reflect the fact that more recent periods are considered more relevant than more distant ones. This modeling choice would be a limitation when we discuss dynamics in the next section, since it requires that the set X depend on T.

[12] See an axiomatization and discussion in Gilboa and Schmeidler [21].

[13] See Akaike [1] and Silverman [40].

more general patterns of observations. Such "higher-level" analogies can also be captured as hypotheses in our model. For instance, the reasoner 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 hypotheses including states in which observations $(i-k+1), ..., i$ are identical to observations $(t-k+1), ..., t$, and so forth.

4.6 Rule-Based Reasoning

Various rules can also be captured by assigning weights to appropriate sets A. For example, consider the rule "it rains whenever the sky is cloudy", where $\omega_x(t) \in \{0, 1\}$ indicates whether the sky is cloudy and $\omega_y(t) \in \{0, 1\}$ indicates whether it rained in observation i. Then the rule can be described by

$$A = \{\omega \in \Omega \,|\omega(t) \neq (1, 0) \quad \forall t\}\,.$$

The rule will be excluded from the summation as soon as a single counter-example is observed. If it has not been excluded, it may or may not apply to the next observation: if we observe clear skies, $\omega_x(t) = 0$, any value $\omega_y(t)$ is compatible with A, and A is irrelevant. However, when the sky is cloudy, the weight $\phi(A)$ will be summed up in $\Phi(1)$ (but not in $\Phi(0)$), supporting the prediction that it is going to rain. More generally, any association rule can be modeled in our framework. Similarly a functional rule, which states that the value of y is a certain function f of the value of x, can also be captured by a hypothesis, such as

$$A = \{\omega \in \Omega \,|\omega_y(t) = f(\omega_x(t)) \quad \forall t\}\,.$$

Holland's [23] 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." At each point of time, each rule has a weight, which 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 apply to the case at hand, and that predict this y.

The prediction part of genetic algorithms is therefore a special case of our model, where the hypotheses are the association rules involved. However, the way that the weights are generated in genetic algorithms does not correspond to a constant model: rules are generated by a partly-random process, including crossover between "parent genes,"mutations, and so forth.

5 Dynamics of Reasoning Methods

In this section, we consider how reasoning changes as a result of evidence. We consider a collection of models, indexed by T, with the value of T growing arbitrarily large. The sets X and Y are assumed to remain the same for all T. In the model with T periods there is a state space Ω_T, with a set of hypotheses $\mathcal{A}_T = 2^{\Omega_T}$. The reasoner uses a model ϕ_T. The sets $\mathcal{B}_T$ and $\mathcal{CB}_T$ are defined as above for each T.

Throughout this section we examine a qualitative reasoner and constant ϕ_T (for each T).

5.1 Bayesian vs. Case-Based Reasoning

Consider models $\{\phi_T\}_T$ that put all their weight on the Bayesian and case-based hypotheses. That is,

$$\phi_T(A) = 0 \qquad \forall A \notin \mathcal{B}_T \cup \mathcal{CB}_T$$

Assume that the weight of the two types of reasoning does not vanish, and normalize the total weight to one. Specifically,

Assumption 1 *For some $\varepsilon > 0$,*

$$\begin{aligned} \phi_T(\mathcal{B}_T), \phi_T(\mathcal{CB}_T) &> \varepsilon \\ \phi_T(\mathcal{B}_T) + \phi_T(\mathcal{CB}_T) &= 1. \end{aligned}$$

Assumption 1 means that the reasoner keeps a certain minimal degree of credence in both modes of reasoning, independent of T.

We can think of the reasoner as allocating the overall weight of credence in a top-down approach: first allocating a weight $\phi_T(\mathcal{B}_T)$ to the Bayesian approach, and the complement weight $\phi_T(\mathcal{CB}_T)$ to the case-based, analogical way of reasoning. Then, the reasoner splits these weights, within each family of hypotheses, among particular hypotheses.

How should the weights be split within each set of hypotheses? We start with the weight of the Bayesian hypotheses, $\phi_T(\mathcal{B}_T)$. If the reasoner knows something about the process she is about to observe, this knowledge should be reflected in her weights, or prior beliefs. In an extreme case, the reasoner might have $\phi_T(\{\omega\}) = 1$ for a particular ω. We are interested in the contrasting case in which the reasoner knows relatively little about the process she is observing, so little that she cannot rule out *any* state. She accordingly assigns a positive weight to each state ω.

How should these prior probabilities be chosen? A simple and common approach is to assume that the reasoner has a uniform prior over the state space. This, however, may be somewhat restrictive. The seek a weaker assumption, requiring only that the probability assigned to any particular state cannot be too much smaller than that assigned to another state. A uniform prior obviously satisfies this, by making every such probability equal, or

$$\frac{\phi_T(\{\omega\})}{\phi_T(\{\omega'\})} = 1$$

for every $\omega, \omega' \in \Omega_T$ and every T. A more general condition would assume that there are bounds on the ratio of the probabilities attached to any two states, i.e., that there exists $\theta \in (0,1)$ such that, for every $\omega, \omega' \in \Omega_T$ and every T,

$$\theta < \frac{\phi_T(\{\omega\})}{\phi_T(\{\omega'\})} < \frac{1}{\theta}. \tag{6}$$

We weaken this assumption further and assume only that the ratio between the probabilities of two states cannot go to infinity (or zero) too fast as T grows. Formally,

Assumption 2 *Openmindedness: There exists a polynomial $P(T)$, such that, for every T and every two states $\omega, \omega' \in \Omega_T$,*

$$\phi_T(\{\omega\}) \leq P(T)\phi_T(\{\omega'\}).$$

Assumption 2 allows for a more general class of beliefs than our first equal-probability or bounded-probability-ratios assumptions. For each T, there exists $\theta = \theta_T \in (0,1)$ satisfying (6), where θ_T is allowed to converge to 0 as $T \to \infty$. The assumption is, however, that this convergence is not too fast, that is, that it is bounded by a polynomial in T. When this assumption holds, we say that the reasoner is *open minded*.

We similarly need some structure on how the reasoner allocates probability among the various case-based hypotheses:

Assumption 3 *There exists a similarity function $s : X \times X \to \mathbb{R}_{++}$ and a decay factor $\beta \in (0, 1]$ such that, for every T, there exists $c_T > 0$ such that, for every $i < t < T$, and every $x, z \in X$,*

$$\phi_T(A_{it,x,z}) = c_T \beta^{t-i} s(x, z).$$

Assumption 3 has two parts. The first is that the similarity function is strictly positive. Second, the weight of a case-based hypothesis $A_{it,x,z}$ is determined by the similarity of the two problems, one with data x in period i and the other with data z in period t. If we do not assume a decaying memory, that is, if $\beta = 1$, Assumption 3 implies stationarity: the weight assigned to the hypothesis that period t will be identical to period i is independent of the time indices i and t. Moreover, these weights can only change proportionately as we vary the time horizon T. Thus we can summarize them by a similarity function, such that the weight of the hypothesis is simply a multiple of this similarity value, namely, $c_T s(x, z)$. If, by contrast, memory does decay, that is, $\beta < 1$, the similarity between observing x at period t and observing z at a preceding period i is the inherent similarity between x and z, $s(x, z)$, multiplied by the decay factor β^{t-i}.

The following result states that, under the assumptions above, if the reasoner has a sufficiently long string of data, then she will become essentially a case-based reasoner.

Theorem 1 *Let Assumptions 1–3 hold. Then for every $\alpha, \delta > 0$ there exists T_0 such that, for every $T > \frac{1}{\alpha} T_0$, every $t \geq \alpha T$, and every history h_t^*,*

$$\frac{\phi_T\left(\mathcal{B}_T \cap R(h_t^*)\right)}{\phi_T\left(\mathcal{CB}_T \cap R(h_t^*)\right)} < \delta.$$

The meaning of this result is that given a sufficiently long horizon, the reasoner will put virtually all of her weight on case-based, rather than on Bayesian hypotheses, for all but a small fraction of initial periods. The basic idea of the proof is simple: the rate of increase of the size of $\mathcal{B}_T$ (with T) is exponential, whereas that of $\mathcal{CB}_T$ is polynomial (in fact, quadratic). As T grows, because t ($\geq \alpha T$) grows with it, the number of states that are compatible with history h_t^* becomes an exponentially small fraction of

the size of $\mathcal{B}_T$, and thus (given open-mindedness) their cumulative weight tends to zero at an exponential rate. By contrast, the number of case-based hypotheses that are compatible with history h_t^* is a linear function of t, and each has a weight that is inversely proportional to T^2. Hence the total weight of the case-based hypotheses tends to zero at a much slower rate.

Note that the only learning that is taking place in our process is the exclusion of refuted hypotheses. As mentioned above, this is the counterpart of Bayesian updating of probabilities. It is interesting that this "Bayesian updating" favors case-based reasoning over Bayesian reasoning.

The Bayesian part of the reasoner's beliefs converges to the truth at an exponential rate as evidence is accumulated (that is, as t grows). That is, the probability of the true state *relative to* the probability of all unrefuted states grows exponentially with t. This increase of the posterior probability of the true state does not result, of course, from any change in the prior probability, but from the exclusion of 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 hypotheses tapers off and leaves the stage to the case-based hypotheses. The very mechanism that makes the posterior probability of truth grow makes the weight of Bayesian reasoning diminish.

Proposition 1 makes a powerful statement: the asymptotic result holds at each and every state of the world. A slightly weaker statement could be made without the open-mindedness requirement, if one were to add probabilistic assumptions to the model. For instance, assume that the process observed is governed by a probability measure λ. Assume that λ and ϕ are independently chosen. Then one can show that, with very high λ-probability, the relative weight of the case-based hypotheses will converge to 1.

It may be worthwhile to highlight the role of each assumption in the derivation of this result. Assumption 1 demands that the reasoner put some a priori weight (ε) on each of the two modes of reasoning. Clearly, a reasoner who a priori assigns zero weight to the case-based mode of reasoning will never switch to it (under our assumption of a constant model ϕ).

Assumption 2 is not only the key to the derivation of our result, but also suggests the scope of applications where the Bayesian approach is reasonable. As such, it deserves a special discussion. Since this discussion is equally relevant to Subsection 5.2, we defer it to Subsection 5.3.

Finally, Assumption 3 states that the case-based weights are governed by

a similarity function that changes proportionally as does the horizon T. An example in which this assumption is violated might be a system of weights according to which a hypothesis $A_{it,x,z}$ is assigned a weight which is exponentially decreasing in t. While logically possible, we find such a pattern neither very likely nor very rational.

Why does the overall weight put on the Bayesian hypotheses shrink faster than that put on the case-based ones? A case-based hypothesis has exponentially many states in it. Should this weight be divided among the relevant states, each would have a small weight as in the Bayesian case. Why is it, then, that the event retains a higher weight than the states, that the whole is more than the some of its part, as it were?

The reason is precisely this: because non-Bayesian hypotheses need not divide the weight among their states, they don't have to say too much, and are thus not so often refuted. Consider the following example. The year is 1925 and different experts are asked about the evolution of the DJIA over the next hundred years. Each expert provides a hypothesis. A Bayesian expert provides a hypothesis that is a singleton. Thus there are many Bayesian experts, and each gets a small a priori weight. Eighty years later, the vast majority of them were proven wrong and only a few are still in the game, resulting in an overall low weight for the Bayesian experts as a group. By contrast, there is an expert who says "year 2008 will be similar to year 1929". This hypothesis is clearly not Bayesian, as it says nothing about the years 1925,...,1928, 1930,...,2007, etc. Saying nothing about these years, the hypothesis is not risking being refuted by the respective observations, and thus more case-based hypotheses are unrefuted by the year 2008.

Now suppose that the reasoner demands that experts specify their hypotheses, that is, that all be Bayesian. This means that the non-Bayesian (case-based) expert who predicted that 2008 would be similar to 1929 is asked to divide the weight of her hypothesis among the exponentially many states that constitute it. But the expert might well say, "I do not have any prediction about 1926 or about 1937. All I said was that there will be similar conditions in 1929 and in 2008." Insisting that the expert further specify the hypothesis may be stretching the limit of her expertise. And if later it is found out that her predictions for the years 1926 or 1937 were falsified, it would be wrong to penalize her original prediction, which did not purport to say anything about these observations.

Put differently, the Bayesian approach is not flexible enough to allow the experts to say "I do not know." It requires that they quantify their beliefs

about all questions. Hence, the Bayesian approach does not allow us to distinguish among experts according to the accuracy of their self-assessment; an expert who knows that she does not know certain probabilities and an expert who wrongly believes that she does know them may end up subscribing to the same assessments. By contrast, other approaches allow for a "don't know" answer, and thus can indirectly give experts credit for knowing the limitations of their knowledge. A non-Bayesian expert may avoid refutation either by making correct predictions, or by knowing when to remain silent. the inclusion of non-Bayesian hypotheses therefore allows the reasoner to judge experts not only the their specific knowledge, but also by their meta-knowledge, namely, the knowledge of what they know and what they do not know.

5.2 Bayesian vs. Rule-Based Reasoning

Rule-based reasoning takes a large variety of forms. One very special type of rules predicts a constant outcome, $y \in Y$, that is, "we will always observe y." As stated, there is but one such rule for each $y \in Y$, and this rule will likely be refuted early on. (To be precise, there is little interest in problems in which one of these rules proves true.) However, a minor variation of these rules are not easily refuted. Consider the rule "As of period i, we will only observe y." If it is the case that only y has been observed in periods $i, (i+1), ..., (t-1)$, it might be tempting to conjecture that only y will be observed from period i on. Such a hypothesis may not be easy to extend to the first i periods in an elegant way. But, whatever the history up to period i, it is always possible that, as of that period, a "regime change" has occurred. For example, people may believe that WWII was the last war in Europe, despite many wars on European soil before WWII. Similarly, lay people as well as experts often predict linear trends, such as "the stock market is on the rise" or "religion is becoming more popular" when these describe recent trends, even if they do not describe less recent history very well.

To capture these types of rules in our model, define, for $t \geq 0$ and $y \in Y$, a hypothesis

$$R_{i,y} = \{\omega \in \Omega \,|\omega_y(t) = y \quad \forall t \geq i\}.$$

Let the set of such rule-based hypotheses be

$$\mathcal{RB}_T = \{R_{i,y} \,|i < T, y \in Y\}.$$

Observe that there are only $|Y|T$ rules in $\mathcal{RB}_T$. We mention again that the term "rule-based reasoning" typically applies to much more general sets of rules, and the present one should only be considered an example.

Even though the number of such rules is only linear in T, there will always be unrefuted rules in this set. Specifically, for $y = \omega_y(t-1)$, the rule $R_{(t-1),y}$, which suggests that the last observation will become the general rule, is both consistent with and relevant at h_t^*. Moreover, for every $z \in Y$, the rule $R_{t,z}$ is also in $\mathcal{RB}_T \cap R(h_t^*)$. Therefore, for every history h_t^*,

$$|\mathcal{RB}_T \cap R(h_t^*)| = |Y| + 1.$$

As in the previous sub-section, we assume that one starts with a model ϕ_T that attaches weight to both Bayesian and rule-based hypotheses:

Assumption 4 *For some $\varepsilon > 0$,*

$$\begin{aligned} \phi_T(\mathcal{B}_T), \phi_T(\mathcal{RB}_T) &> \varepsilon \\ \phi_T(\mathcal{B}_T) + \phi_T(\mathcal{RB}_T) &= 1. \end{aligned}$$

Hence, the reasoner focuses only on Bayesian and rule-based hypotheses, allowing each class of rules an overall minimal weight $\varepsilon > 0$ independently of T.

How should the weights be assigned within each class? We continue to assume that the reasoner is open minded with respect to the Bayesian hypotheses, that is, that Assumption 2 holds. As in Proposition 1, this implies that the total weight assigned to relevant Bayesian hypotheses decreases exponentially with T. As for the rule-based hypotheses, we start with the benchmark assumption that the weight is allocated to them uniformly, that is,

$$\phi_T\left(R_{i,y}\right) = \frac{1}{|Y|T}.$$

Such an assumption would clearly imply that the reasoner converges to rule-based reasoning, since

$$\phi_T\left(\mathcal{RB}_T \cap R(h_t^*)\right) = \frac{|Y|+1}{|Y|T} > \frac{1}{T}$$

for all h_t^*, and a result comparable to Proposition 1 follows.

However, the uniform distribution is not a very reasonable assumption in this case. Suppose that the observations in periods $k \in \{i, ..., t-1\}$ have been

constant, that is, $\omega_y(k) = y$ for some y. A uniform distribution of weight would put all rules $R_{k,y}$ on equal footing, despite the fact that $R_{i,y}$ has made correct predictions over the past $t - i$ periods, whereas $R_{t,y}$ remained silent on them. Similarly, after observing y in period 0, it seems that the rule $R_{0,y}$, which has been proven right at least once, should have more weight than rule $R_{1,z}$, which makes the z observation out of the blue.

It therefore appears more plausible to assume that the weight of $R_{i,y}$ decreases with i. In particular:

Assumption 5 *Let either*

$$\phi_T(R_{i,y}) = \rho^i$$

for $\rho \in (0,1)$ *or*

$$\phi_T(R_{i,y}) = \frac{1}{i^\nu}$$

for $\nu > 1$.

The arguments analogous to those used to prove Proposition 1 give:

Proposition 2 *Let Assumptions 2, 4 and 5 hold. Then for every* $\alpha, \delta > 0$ *there exists* T_0 *such that, for every* $T > \frac{1}{\alpha}T_0$, *every* $t \geq \alpha T$, *and every history* h_t^*,

$$\frac{\phi_T(\mathcal{B}_T \cap R(h_t^*))}{\phi_T(\mathcal{RB}_T \cap R(h_t^*))} < \delta.$$

Hence, when the horizon is long, the weight of the Bayesian hypotheses will be negligible compared to the weight of the rule-based hypotheses in all but a handful of initial periods.

5.3 When is Bayesianism Reasonable?

The discussion in the previous two sub-sections suggests that there are circumstances in which Bayesianism is not a very reasonable way of reasoning about the world. In particular, we observed that under certain assumptions the Bayesian way of reasoning is guaranteed to die out and leave the stage to others. Clearly, these results depend on the underlying assumptions, and there are circumstances under which Bayesian reasoning will remain useful

and even dominant. To consider a trivial example, consider a reasoner who is a devout Bayesian, satisfying

$$\phi_T(\mathcal{B}_T) = 1$$

for all T. Such a reasoner will obviously remain Bayesian in the face of whatever evidence she gathers. In these cases, the result that Bayesian reasoning is discarded does not hold because Assumptions 1 and 4 do not hold.

This example is not only trivial but also extreme. Indeed, under the relevant assumption (1 or 4), should one allow for the smallest doubt, and assign a positive weight either to a case-based or to a rule-based mode of reasoning, then the Bayesian way of thinking is driven out by its competitors. Interpreting the weights as subjective probabilities regarding the theory that actually governs the data generating process, it suffices that a very small probability is assigned to the non-Bayesian ways of thinking, to shrink the weight put on the Bayesian approach, as a result of a pseudo-Bayesian update.

However, Propositions 1 and 2 crucially depend on Assumption 2, to which all eyes now turn. Two examples in which this assumption does *not* hold will be helpful. First, Assumption 2 is obviously violated if the reasoner believes that she knows the true state of the world, say, if for some ω_T, $\phi_T(\{\omega\}) = 1 - \varepsilon$ for all T. If, on top of the above, the reasoner is also correct in her focus on state ω_T, that is, at state ω_T, the posterior probability attached to Bayesian hypotheses will never dip below $1 - \varepsilon$. In other words, if the reasoner believes she knows the truth, and happens to be right, her Bayesian beliefs will remain dominant.

A slightly less trivial example is the following. Consider, for concreteness, Bayesian and case-based reasoning alone, that is, suppose that Assumption 1 holds. For simplicity, let $X = \{0\}$ and $Y = \{0, 1\}$. That is, all periods have the same observable features, and they only differ in the binary variable the reasoner is trying to predict. Suppose that the reasoner believes that she observes a cyclical process. Formally, for $1 \leq k \leq T$, let $\omega^k \in \Omega_T$ be defined by

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

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

Let the reasoner's beliefs satisfy

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

and

$$\phi_T(\{\omega\}) = 0$$

for every $\omega \notin \left\{\omega^k \,|1 \leq k \leq T\right\}$. Thus, the reasoner splits all the weight of the Bayesian hypotheses among the k hypotheses $\left\{\omega^k\right\}$ and leaves no weight to the other Bayesian beliefs.[14]

Next suppose that the reasoner is right in her belief that the process is cyclical (starting with a sequence of 0's). That is, focus on one of the states ω^k. Clearly, once we get to period $t = k$, all the Bayesian hypotheses $\left\{\omega^{k'}\right\}$ for $k' \neq k$ are refuted. By contrast, the hypothesis $\left\{\omega^k\right\}$ is not refuted at any t. Consequently, at ω^k, for every $t \geq k$, the total weight of the Bayesian hypotheses remains $\frac{1-\varepsilon}{2^k}$. This weight does not depend on T. By contrast, the total weight of the case-based hypotheses decreases with T, resulting in the Bayesian mode of reasoning becoming the dominant one. Clearly, this will only be true at the states ω^k. At other states the converse result holds, because all Bayesian hypotheses will be refuted and case-based reasoning will be the only remaining mode of reasoning.

Two main assumptions are needed for the success of the Bayesian approach in these examples. First, the reasoner has to believe that some states are much more likely than others. Second, she has to be right. If the reasoner knows the data generating process up to certain parameters (such as the cycle length, k, in the last example), then Bayesian beliefs allow learning and need not be driven out by other forms of reasoning. Indeed, if a Bayesian reasoner knows the data generating process up to a fixed number of parameters (that does not grow with T), she has no reason to choose beliefs that conform to Assumption 2. In these applications the Bayesian approach appears very reasonable. If the data generating process is known up to the specification of a few parameters, and even more so if these parameters can assume only finitely many values (or belong to a compact space), it makes sense to specify a prior over the set of parameters and to update that prior as observations are gathered. This type of Bayesian reasoning is successful because the set of parameters does not increase with the number of observations. Put differently, the agent is not learning the state space Ω_T, which grows with T, but the parameter space, which is fixed. This type of learning can also be viewed as rule-based.

By contrast, assume that the reasoner observes a process about which

[14] Observe that these Bayesian beliefs can also be described as rule-based beliefs. As we argue below, this is not a coincidence.

nothing is known a priori. She may be interested in stock market behavior or in the eruption of wars. In either case there is no claim that the data generating process is known up to a finite set of parameters. In these examples the reasoner has to form her beliefs over the entire state space and not over a parameter space. Moreover, the size of the state space increases with T, at an exponential rate. Hence it seems a priori harder to learn the process. Correspondingly, we find Assumption 2 rather natural for such examples. In fact, one can argue that it is irrational to violate it, as such a violation suggests that the reasoner believes she knows about the process more than she actually does.

The distinction between a fixed parameter space and the entire, all-encompassing state space also appears when one contrasts applications of the Bayesian approach in economics with Bayesian applications in statistics, computer science, and even the philosophy of science. In the latter, the object of uncertainty is typically a space of parameters, conjectures, or theories, that does not grow with the accumulation of data. In contrast, economic theory has adopted the "Harsanyi Doctrine", suggesting that rationality necessitates a Bayesian approach to the Grand State Space, in which a state "resolves all uncertainty."[15]

Our results should therefore not be viewed as a critique of the Bayesian approach as employed in statistics, computer science, and similar settings. These applications tend to apply Bayesian beliefs to a fixed parameter space, and when we derive the corresponding implicit Bayesian beliefs over the state space, Assumption 2 is violated. Intuitively, the reason is that in these applications the reasoner knows the basic structure of the data generating process. In contrast, when no such knowledge exists, our results suggest that, under a mild assumption, the Bayesian approach will be discarded as more observations are gathered.

5.4 Case-Based vs. Rule-Based Reasoning

The previous sections argued that both case-based reasoning and rule-based reasoning may come to supplant Bayesian reasoning, unless the reasoner's

[15]The tern "resolves all uncertainty" goes back to Savage [36]. However, there are numerous references that indicate that Savage did not intend his approach to apply to states that describe history in its entirety. See also Gilboa, Postlewaite, and Schmeidler [17] on the distinction between the way the Bayesian approach is applied in economic theory and in the other disciplines.

prior belief already encompasses a great deal of information about the world. When faced with complicated phenomena, which are not repeated in the same manner but which may be causally intertwined, reasoners will tend to put more weight on case-based and rule-based reasoning methods than on the Bayesian one. But what determines (or should determine) the trade-off between case-based reasoning and rule-based reasoning? When should reasoners stick to simple analogies and when should they engage in theorizing?

Both modes of reasoning are essential for effective learning, and a healthy balance should be maintained between them. Case-based reasoning may miss very obvious trends. For example, if a process y_t grows linearly in t, an obvious rule begs to be detected, but a case-based reasoner will refuse to recognize it, and will keep predicting that y_t will be some average of past values $y_0, ..., y_{t-1}$. By contrast, rule-based reasoning is prone to over-fitting.[16] We devote this section to two examples of over-fitting, where case-based reasoning may mitigate the problem and improve predictions.

5.4.1 Example 1: Theory cycles

Assume that there are no predicting variables, or equivalently, that the value of x is constant: $|X| = 1$. Let $Y = \{0, 1\}$ and assume that y_t are i.i.d., where $y_t = 1$ with probability p.

Consider the set of rules defined above,

$$\mathcal{RB}_T = \{R_{i,y} \,|i < T, y \in Y\},$$

where

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

for $t \geq 0$ 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. There are $2T$ hypotheses in $\mathcal{RB}_T$.

The case-based hypotheses are rather simple: since there are no x values to consider, the hypotheses are simply

$$A_{it} = \{\omega \in \Omega \,|\omega_y(i) = \omega_y(t)\}.$$

[16]Rule-based and the case-based approaches may be viewed as analogous to chartists and fundamentalists' views of stock market data: the chartists tend to look for trends, running the risk of findng trends when they do not exist, while the fundamentalists pay more attention to past data, and may fall prey to the opposite bias.

Thus, the set of all case-based hypotheses is

$$\mathcal{CB}_T = \{A_{it} \,|i < t \leq T\}$$

and it contains $T(T-1)/2$ hypotheses.

Assume that, for $0 < c < 1$

$$\begin{aligned} \phi_T(\mathcal{CB}_T) &= c \\ \phi_T(\mathcal{RB}'_T) &= 1 - c, \end{aligned}$$

and, for simplicity, that the weights within each class of hypotheses are uniformly distributed. Thus, for $t \geq 0$ and $y \in \{0, 1\}$,

$$\phi_T(R_{i,y}) = \frac{1-c}{2T}$$

and for $i < t \leq T$,

$$\phi_T(A_{i,t}) = \frac{2c}{T(T-1)}.$$

Next, let us consider histories h_t^* ending with k ($< t$) 1's, that is, $y_i = 1$ for $t - k \leq i < t$, but $y_{t-k-1} = 0$. For each $t - k \leq i \leq t$, the rule $R_{i,1}$ is relevant at h_t^* and predicts $y_t = 1$. There is one more rule that is relevant at h_t^*, and this is $R_{t,0}$, which predicts 0. Overall the rule-based prediction contributes $\frac{(k+1)(1-c)}{2T}$ to the prediction 1, and $\frac{1-c}{2T}$ to the prediction 0. The case-base prediction splits the weight $\frac{2tc}{T(T-1)}$ between 0 and 1 proportionally to the average

$$\bar{y}_{t-1} = \frac{1}{t} \sum_{i=0}^{t-1} y_i.$$

Assume that T is large. For small values of t, the overall weight of the case-based prediction is $O(T^{-2})$ and it therefore does not significantly change the rule-based prediction. Thus, even if k is small, the reasoner will predict $y_t = 1$. This phenomenon means that the reasoner is a little too quick to find trends in the data: a few observations of the value 1 suffice for her to theorize that "from now on, we'll observe only 1."We may view this phenomenon as a type of "overfitting": the reasoner finds a theory that matches the data perfectly, but it only discusses the most recent observations.

Next consider large values of t, say, $t = \alpha T$ for $\alpha \in (0, 1)$. The total weight of the case-based hypotheses is now

$$\frac{2tc}{T(T-1)} = \frac{2\alpha c}{(T-1)}$$

and it is of the same order of magnitude as the total weight of the rule-based predictions,

$$\frac{(k+1)(1-c)}{2T}.$$

The ratio of the above is

$$\frac{\frac{2\alpha c}{(T-1)}}{\frac{(k+1)(1-c)}{2T}} = 4\alpha \frac{T}{T-1}\frac{c}{1-c}\frac{1}{k+1}.$$

It follows that for low values of c, reflecting a tendency to theorize, the reasoner will tend to over-generalize and find patterns even in data that are in fact random. By contrast, high values of c, associated with a tendency to rely on experience, will reduce the chance of overfitting the data and over-generalizing trends, at the cost of ignoring trends when they actually exist.

Intermediate values of c would result in rule-based reasoning being dominant for large k (relative to c and α), and case-based reasoning taking over when k is small. In other words: when recent history is suggestive of a simple rule (a large number of observations of 0 or of 1), the reasoner adopts the rule "recent observations will continue forever."When recent history is more spotty, and no simple rule explains it, the reasoner 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, for a large T 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 reasoners, and at some point it will collapse. When it does collapse, confusion may lead reasoners 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, the reasoner will tend to prefer it over case-based reasoning. But when all simple theories are refuted, the reasoner 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. On the bright side, they can never be refuted. They are always there, waiting faithfully for the reasoner, who would devote more attention to them when her heroic attempts to understand the process fail.

5.4.2 Example 2: Multiple Predictors

Assume that there are m binary predictors, that is, that $X = \{0,1\}^m$ and denote $x = (x(1), ..., x(m))$. Assume also that $Y = \{0,1\}$. Consider a reasoner who employs only case-based and rule-based reasoning. For simplicity, consider only simple rules suggesting that a single predictor equals the predicted variable y. Formally, for $i < T$ and $j \leq m$, let

$$R'_{i,j} = \left\{ \omega \in \Omega \left| \omega_y(t) = \omega_{x(j)}(t) \quad \forall t \geq i \right. \right\}$$

and

$$\mathcal{RB}'_T = \left\{ R_{i,j} \left| i < T, j \leq m \right. \right\},$$

and assume that

$$\begin{aligned} \phi_T(\mathcal{CB}_T), \phi_T(\mathcal{RB}'_T) &> \varepsilon \\ \phi_T(\mathcal{CB}_T) + \phi_T(\mathcal{RB}'_T) &= 1. \end{aligned}$$

Regarding case-based reasoning, assume that the similarity between two cases depends on the number of predictors that are identical between them:

$$s\left((x(1), ..., x(m)), (z(1), ..., z(m)) \right) = \left(\sum_{j=1}^{m} \mathbf{1}_{\{x(j)=z(j)\}} \right)^{\lambda}.$$

Thus, when $\lambda = 0$ the reasoner ignores the x values and bases her prediction on empirical frequencies of the y values, and when $\lambda \to \infty$, she tends to take into account only the past cases that were completely identical to the case at hand.

To highlight the overfitting problem, assume that the x variables do not contain any relevant information, that is, that y is independent of x. More generally, one may assume that several x variables have already been fruitfully used for prediction, and y represents only the error term. However,

m is large because there is a vast set of additional variables that are readily observable. The overfitting problem arises because, among these many variables, there are likely to be some whose recent pattern of observations perfectly matches that of y. This would be reflected in the fact that the reasoner would find, among the rules in $\mathcal{RB}'_T$, some rules that cannot be refuted by recent observations. Specifically, for every $k < T$ there is a large enough m such that, with a high probability, one of the m variables, x_j, has the same last k observations as does y. The reasoner will then be tempted to predict y based on the current value of x_j, while, in reality, x_j has no informational value.

By contrast, case-based reasoning is not prone to the same type of overfitting. Moreover, a more permissive similarity function, with a smaller value of λ, will be more robust to overfitting.

Clearly, a larger a priori weight on case-based reasoning, or a lower value for λ, would increase the weight of reasoning by analogies also when rule-based reasoning is needed. If it so happens that y is indeed a function of a few predictors, case-based reasoning might becloud this relationship. This suggests that in different set-ups one may find the optimal mix of case-based and rule-based reasoning, allowing the reasoner to find patterns when they exist, but to rely on simple analogies when the observed patterns are spurious.

6 Discussion

6.1 Belief Functions and Choquet Capacities

A model ϕ assigns non-negative weights to subsets of a state space, Ω. Hence, for every hypothesis $A \subset \Omega$ we may define

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

which is a Choquet [8] capacity, and, in fact, also a Belief Function in the terms of Dempster [11] and Shafer [39]. Choquet capacities have been introduced into decision theory by Schmeidler [38], who axiomatized them in the context of maximization of Choquet expected utility.

The basic intuition behind belief functions as well as some interpretations of Choquet capacities suggest that one may believe in an event (or a hypothesis in our case) without being able to say how the weight of belief should be

split among the states in the event. However, to the best of our knowledge, this interpretation of Choquet capacities has not been applied to modeling case-based or rule-based reasoning.

6.2 Methods for Generating Hypotheses

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

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 hypotheses, with the understanding that within each set the most successful hypotheses is selected for prediction, and, accordingly, the degree of success of the set is judged by the accuracy of this most successful hypothesis. The following example illustrates.

Suppose that the reasoner 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 reasoner 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 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 hypotheses – all hypotheses that predict cycles as above, parametrized by k. If, indeed, 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 hypotheses it generates differ from one dataset to the next.

6.3 Between Cases and Rules

In this paper we discuss the simplest form of case-based reasoning, and several examples of rather primitive rule-based reasoning. These examples suggest that the two modes of reasoning are very different and easily distinguishable. However, there are other examples in which it is harder to draw this distinction.

Suppose, for example, that Ann, Bob, and Christine are trying to predict the outcome of a US presidential election. Ann compares each candidate to each past president, sums up the similarity values, and chooses the candidate with the highest sum. This is clearly a case-based method. Bob, in contrast, believes that the tallest candidate always wins the election. This is a simple rule, and Bob's method is rule-based. Christine believes that the winner will be the candidate who is most similar to JFK. How would we classify Christine's reasoning? On the one hand, it is case-based, as it relies on similarity to a past case. On the other hand, this is a simple rule for prediction. Like Bob, Christine predicts that the winner will be a maximizer of a simple function, defined by finitely many parameters, and hence deserves to be dubbed rule-based.

While this example seems to be both rule-based and case-based, one can think of other examples that are neither. Assume that each observation consists of $x_i, y_i \in \mathbb{R}$ and consider the obvious extension of our model to this set-up (with continuous variables). Given history h_t^*, for each $i < j < t$ one can define y_{ijt} to be the linear extrapolation of y_i, y_j to period t. Assume that the reasoner judges the similarity of x_t to x_i and to x_j, $s(x_i, x_j, x_t)$ and computes a prediction that is the (s) similarity-weighted average of all $(y_{ijt})_{i<j<t}$. Is this a case-based or a rule-based method? On the one hand, the linear extrapolation from (x_i, y_i) and (x_j, y_j) to (x_t, y_{ijt}) is based on a simple linear rule. On the other hand, this reasoning is based on a similarity-weighted aggregation just as our examples of case-based reasoning. In fact, it can be viewed as a slight generalization: the case-based reasoning discussed above aggregates over constant functions (each predicting that y_t be equal to a certain y_i), while the current example aggregates over linear functions (each predicting that y_t be on a line defined by certain y_i,y_j).

Our model can capture these examples that are between case-based and rule-based reasoning. Indeed, it can capture other modes of reasoning as well. As long as the reasoner makes some non-trivial predictions, she can be viewed as stating certain hypotheses: any non-trivial prediction has an

extension, consisting of all states at which it holds.

6.4 Generalization to Probabilistic Theories

Our model can be extended to capture probabilistic theories by allowing the object of knowledge to be probability distributions over the states of the world, Ω, rather than the states themselves. Ex-post, a particular state $\omega \in \Omega$ is revealed, but ex-ante only a distribution over the states is considered knowable. This is a generalization of the model presented above, when restricted to Bayesian hypotheses. Specifically, the Bayesian hypotheses are the deterministic distributions (each assigning probability 1 to a particular state ω). More generally, one may wish to restrict attention to other classes of knowable distributions, perhaps none of which is deterministic.

When considering only specific distributions over Ω, our model generalizes in a very natural manner: each distribution f has an a priori weight $\phi(\{f\})$. Given a history h_t^*, the theory 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 $\phi(\{f\})$ leads to a natural measure of the belief in theory f following history h_t^*. Indeed, this is precisely the result of a Bayesian update over theories given a history.

This model calls for two extensions. The first involves the inclusion of subsets of distributions, and the second for distributions over subsets (and subsets of such distributions). We start with the former.

Consider a hypothesis B that is a set of distributions, rather than a singleton $\{f\}$. A natural extension of the model suggested in the previous subsection leads to the maximum likelihood principle: of all the specific distributions in a set of distributions, choose one that maximizes the likelihood function and use it for the evaluation of the set as a whole. It is as if the family of distributions is credited with the success of its most prominent member. If this maximizer is unique, it should also represent its family in generating predictions. Otherwise, for each possible observation we may select one of the maximizers of the likelihood function that also maximizes the probability of that observation.

Formally, let F be the set of probability vectors over Ω. Let $\mathcal{B}$ be a set of hypotheses, such that every $B \in \mathcal{B}$ is a subset of F. It is natural to constrain $\mathcal{B}$ to include only closed and measurable subsets of F. Let $\phi : \mathcal{B} \to \mathbb{R}_+$ have

finite support. Given a history $h \in \mathcal{H}_t$, and a theory $f \in F$, let $f(h)$ be the probability of h given f. For a history $h_t^* \in H_t^*$, $f(h_t^*)$ is the likelihood of f given h_t^*. For a history $h_{t+1} = (h_t^*, y) \in H_{t+1}$, $f(h_{t+1})$ is the likelihood of f given h_t^*, multiplied by the conditional probability of y given h_t^* according to f. We can now define

$$\phi\left(B|h_t^*\right) = \phi(B) \max_{f \in B} f(h_t^*)$$

and

$$B\left(h_t^*\right) = \arg\max_{f \in B} f(h_t^*).$$

Next, for every $y \in Y$, define $h_{t+1} = (h_t^*, y)$ and

$$\Phi(y|h_t^*) = \sum_{B \in \mathcal{B}} \phi(B) \max_{f \in B(h_t^*)} f((h_t^*, y)).$$

This formulation allows us to capture a statement such as "I would use regression analysis now, because it has performed well in the past."This statement does not commit to a particular regression model. The regression model used in the current problem may have very different parameters than the regression models used in the past. However, since the weight $\phi(B)$ is assigned to the class of regression models, rather than to a particular member thereof, if the maximum-likelihood regression model in the past performed well, the reasoner will tend to use the maximum-likelihood regression model in the current problem, even if it is a different model.

Observe that in the above formulation there is no counterpart to the notion of relevance. The reason is that we are only dealing with distributions over states, and thus each distribution is "relevant" at each history it is consistent with: every f splits the conditional probability of 1 in a way that says something non-trivial about the y to be observed. More generally, one may consider not only distributions over states in Ω, but also over hypotheses, namely, members of 2^Ω. With this, richer, language, one can capture a rule such as "Use regression analysis whenever $x_t \in X_0$" – which is evaluated by the maximum likelihood model in every observation i with $x_i \in X_0$, but whose evaluation is unaffected by periods i where $x_i \notin X_0$.

6.5 Single-Hypothesis Predictions

We have so far discussed aggregated prediction. It is useful to contrast it with an alternative approach, according to which a single hypothesis is se-

lected for each history, with that hypothesis then determining the reasoner's prediction.[17] Formally, given history h_t^*, let there be a preference relation over hypotheses

$$\succcurlyeq_{h_t^*} \subset \mathcal{A} \times \mathcal{A}$$

be a weak order over hypotheses, defined for each history h_t^*. Define the induced *single-hypothesis* relation over predictions to be $\succsim^{\infty}_{h_t^*} \subset Y \times Y$ as follows:

Definition 2 *For every history h_t^* and $y \in Y$, y is* likely *if there exists a hypothesis $A \in R(h_t^*) \cap C((h_t^*, y))$, such that $A \succcurlyeq_{h_t^*} A'$ for all $A' \in R(h_t^*) \cap C(h_t^*)$. For every $y, y' \in Y$, $y \succsim^{\infty}_{h_t^*} y'$ if y is likely, or if neither y nor y' are likely.*

That is, the prediction y is deemed "likely," or "as likely as any other prediction,"if and only if there exists a hypothesis that is maximal under $\succcurlyeq_{h_t^*}$ and that endorses y as the next prediction. This definition is simplified if $\succcurlyeq_{h_t^*}$ is anti-symmetric (i.e., A and A' are indifferent under $\succcurlyeq_{h_t^*}$ only if $A = A'$), so that there is a unique maximum of $\succcurlyeq_{h_t^*}$. In any event, the relation $\succsim^{\infty}_{h_t^*}$ has at most two equivalence classes.

The single-hypothesis prediction relation can be viewed as a limiting case of aggregated prediction. Consider an aggregated prediction generated by the model ϕ. For $\gamma > 0$, consider the model ϕ^γ, defined pointwise: $\phi^\gamma(A) = [\phi(A)]^\gamma$. Let Φ^γ be the aggregation corresponding to ϕ^γ using (1)–(2). If we let γ tend to ∞, the aggregated prediction provided by Φ^γ approaches a single-hypothesis prediction. To make this precise, given a model $\phi : \mathcal{A} \to \mathbb{R}_+$, define $\succcurlyeq_{h_t^*} \subset \mathcal{A} \times \mathcal{A}$ by

$$\begin{array}{lrcl} A \succcurlyeq_{h_t^*} A' & \text{if} & & \\ \quad \text{(i)} & A, A' & \in & R(h_t^*) \text{ and } \phi(A) \geq \phi(A') \\ \text{or (ii)} & A & \in & R(h_t^*) \text{ and } A' \notin C(h_t^*) \\ \text{or (iii)} & A, A' & \notin & R(h_t^*) \text{ and } \phi(A) \geq \phi(A') \end{array}$$

We say in this case that ϕ represents $\succcurlyeq$. Notice that if ϕ represents $\succcurlyeq$, then so does ϕ^γ for any $\gamma > 0$.

[17] Gigerenzer and his collaborators (e.g., [15, Chapters 4–8]) have championed "one-reason" (or, in our terms, single hypothesis) decision making as a description of behavior, and as an appropriate prescription in many circumstances.

The (omitted) proof of the following is then straightforward:

Proposition 3 *Let the antisymmetric relation $\succcurlyeq_{h_t^*} \subset \mathcal{A} \times \mathcal{A}$ induce the singe-hypothesis prediction $\succsim_{h_t^*}^{\infty}$, and let $\phi : \mathcal{A} \to \mathbb{R}_+$ represent $\succcurlyeq_{h_t^*}$. Then there exists $\bar{\gamma}$ such that, for all $\gamma > \bar{\gamma}$, the corresponding aggregate prediction $\succsim^{\gamma}$ generated by ϕ^{γ} via (2) satisfies*

$$y \succ_{h_t^*}^{\infty} y' \implies y \succ_{h_t^*}^{\gamma} y'.$$

In other words, increasing the relative differences between high and low values of ϕ results in an ordering over predictions in Y that almost coincides with the single-hypothesis prediction. The "almost" qualification reflects the fact that the predictions might be equivalent according to the single-hypothesis prediction, while the aggregated prediction discerns between them even for high values of γ. That is, it is possible that $y \sim_{h_t^*}^{\infty} y'$ but $y \succ_{h_t^*}^{\gamma} y'$ for every $\gamma > 0$.

We now show that particular choices of the relation $\succcurlyeq_{h_t^*}$ yield familiar but disparate reasoning techniques.

6.5.1 Simplicism

Consider a reasoner who, among all equally-general and equally-correct theories, selects the one that appears simplest. The preference for simplicity as a guiding principle in the choice of theories was famously suggested by William of Occam in the 14th century. In modern terms, his claim was normative in spirit: he argued that one *should* select the simplest theory. The claim that this is what people *tend* to do, that is, a descriptive interpretation of the preference for simplicity, was suggested by Wittgenstein [43], who stated that the process of induction consists in choosing the simplest theory that conforms to observations. This approach has sometimes been dubbed "simplicism."[18]

Simplicism can apply to a variety of theories. For example, we may consider only hypotheses that are fully specified, i.e., hypotheses A that contain but a single state of the world each: $|A| = 1$. The aggregated prediction in this case led to Bayesian inference (Section 4.4). By contrast, the single-hypothesis prediction results in a simplicistic approach: after each

[18] See Solomonoff [41], who suggested to couple the notion of simplicism with Kolmogorov complexity to yield a theory of philosophy of science. Gilboa and Samuelson [18] discuss the optimal selection of the relation $\succsim_{h_t^*}$ in this context.

history h_t^*, the reasoner selects the simplest state of the world ω that is consistent with h_t^*.

Formally, we may consider a relation over hypotheses $\succcurlyeq_{h_t^*}$ such that

$$A \succ_{h_t^*} B$$

whenever

$$|A| = 1 \qquad |B| > 1$$

For every history h_t^* there are states of the world ω that are consistent with it, and thus the reasoner will never select a hypothesis B with $|B| > 1$. Clearly, the same principle can be applied more generally to hypotheses that are not necessarily singletons.

Conversely, a reasoner who always selects a single hypothesis, according to some $\succcurlyeq_{h_t^*}$, to generate predictions can be viewed as a simplicity-seeking reasoner, if we interpret $\succcurlyeq_{h_t^*}$ as a simplicity relation.

6.5.2 Nearest Neighbor Prediction

Consider a relation over hypotheses $\succcurlyeq_{h_t^*}$ that places at the top the case-based hypotheses $\{A_{it,x,z}\}_{i,t,x,z}$ as defined in Subsection 4.5. Among these hypotheses, assume that $\succcurlyeq_{h_t^*}$ agrees with a similarity function $s : X \times X \to \mathbb{R}_+$ (taking $\beta = 1$). That is,

$$A_{it,x,z} \succcurlyeq_{h_t^*} A_{it,x',z} \qquad \iff \qquad s(x,z) \geq s(x',z)$$

The resulting single-hypothesis prediction is obtained by choosing the past case that is most similar to the present one, and making a prediction that the present case will result in the same y value as the most similar one in the past. This is equivalent to a (single) nearest-neighbor approach, for a distance function that is the inverse of the similarity function. Such approaches were suggested by Fix and Hodges ([12, 13]) and Cover and Hart [9] and have been applied in a large variety of problems.

Nearest neighbor (NN) approaches have been generalized to rely on more than one sample point. Specifically, k-NN uses the k nearest neighbors, with a certain aggregation rule among them, such as majority vote. The number k may increase with the size of the database, but it should remain small relative to the latter. k-NN approaches are a hybrid between the single-hypothesis and aggregated methods discussed here. Like the latter, they involve some form of aggregation; like the former, they discard the relatively less weighty

hypotheses. Such a hybrid approach is particularly attractive in actual applications, where reducing the relevant database (for each prediction) has computational advantages. Correspondingly, such a hybrid approach can be applied to general hypotheses as discussed here.

6.5.3 Comparison

We can summarize some of the special cases of our model of induction as follows:

Prediction Method	Size of typical hypothesis, $\lvert A\rvert$: 1	Size of typical hypothesis, $\lvert A\rvert$: $(\lvert X\rvert \cdot \lvert Y\rvert)^{n-2}$
Aggregated	Bayesian	Case-Based
Single-Hypothesis	Simplicism	Nearest Neighbor

7 Appendix: Proofs

7.1 Proof of Proposition 1

Fix a quantitative prediction $\Pr(y|h_t^*)$. The easiest route to construction a model generating this behavior is to construct a Bayesian model. Fix a sequence $(x_0, \ldots, x_T)$ and fix a state ω with $\omega_x(t) = x_t$ for $t = 0, \ldots T$. Then letting

$$\phi(\{\omega\}) = \Pr(\omega_y(0)|h_0^*) \Pr(\omega_y(1)|h_1^*) \cdots \Pr(\omega_y(T)|h_T^*)$$

gives the result. ■

7.2 Proof of Theorem 1

Let Assumptions 1 and 2 hold. Let there be given $\delta > 0$. We need to show that there exists T_0 such that, for every $T > T_0$, every $t \geq T_0$, and every history h_t^*,

$$\frac{\phi_T\left(\mathcal{B}_T \cap R(h_t^*)\right)}{\phi_T\left(\mathcal{CB}_T \cap R(h_t^*)\right)} < \delta.$$

We first bound the numerator from above, using Assumption 2. Then we bound the denominator from below, using Assumption 3.

We start by showing that, because the ratio of weights assigned to specific states (hypotheses in $\mathcal{B}_T$) is bounded by a polynomial, the weight of each particular state is bounded by the polynomial divided by an exponential function of T.

Consider a state ω. If $\phi_T(\{\omega\}) > \eta$, then, since for every $A, B \in \mathcal{B}_T$, $\phi_T(A) \leq P(T)\phi_T(B)$, for every ω',

$$\phi_T(\{\omega'\}) \geq \frac{\phi_T(\{\omega\})}{P(T)} > \frac{\eta}{P(T)}$$

Observe that $|\Omega| = d^T$ for $d = |X||Y|$. Hence

$$\phi_T(\mathcal{B}_T) > \frac{d^T \eta}{P(T)}$$

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

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

Hence, for every ω,

$$\phi_T(\{\omega'\}) \leq \frac{P(T)}{d^T}.$$

Next, we wish to show that the weight of all the states that are consistent with a given history h_t^* has to be relatively small. Observe that, because $t < T$, there are exponentially many states that are consistent with h_t^*. Indeed, if t were fixed, their total weight need not converge to zero. However, we assume that t is at least a fixed proportion, α, of T. This implies that the total weight of all states consistent with h_t^* decreases exponentially with T. Specifically, for every h_t^*,

$$|\mathcal{B}_T \cap R(h_t^*)| = d^{T-t}$$

and it follows that

$$\phi_T\left(\mathcal{B}_T \cap R(h_t^*)\right) \leq d^{T-t}\frac{P(T)}{d^T} = \frac{P(T)}{d^t}$$

and since $t \geq \alpha T$,

$$\phi_T\left(\mathcal{B}_T \cap R(h_t^*)\right) \leq \frac{P(T)}{d^{\alpha T}} = \frac{P(T)}{\left(d^\alpha\right)^T} \tag{8}$$

We now turn to bound the denominator, showing that the total weight of all case-based hypotheses that are consistent with h_t^* decreases only at a polynomial rate (as $T \to \infty$). We start with the case of no decay, $\beta = 1$. In this case

$$|\mathcal{CB}_T \cap R(h_t^*)| = t$$

The number of all case-based hypotheses is

$$|\mathcal{CB}_T| = \frac{T(T-1)}{2}|X|^2$$

where, for every $x, z \in X$, there are $\frac{T(T-1)}{2}$ hypotheses of the form $A_{it,x,z}$. Each has a weight

$$\phi_T(A_{it,x,z}) = c_T s(x,z)$$

and thus

$$\phi_T\left(\mathcal{CB}_T\right) = c_T\frac{T(T-1)}{2}\sum_{x,z\in X} s(x,z).$$

Denote

$$\begin{aligned}\hat{s} &= \max_{x,z\in X} s(x,z)\\ \check{s} &= \min_{x,z\in X} s(x,z) > 0\end{aligned}$$

On the one hand,

$$\phi_T\left(\mathcal{CB}_T\right) \leq c_T \frac{T(T-1)}{2} \left|X\right|^2 \hat{s}$$

and since $\phi_T\left(\mathcal{CB}_T\right) > \varepsilon$,

$$c_T \frac{T(T-1)}{2} \left|X\right|^2 \hat{s} > \varepsilon$$

or

$$c_T > \frac{2\varepsilon}{T(T-1)\left|X\right|^2 \hat{s}}.$$

On the other hand,

$$\begin{aligned}\phi_T\left(\mathcal{CB}_T \cap R(h_t^*)\right) &\geq tc_T\check{s} > \\ \frac{2t\check{s}\varepsilon}{T(T-1)\left|X\right|^2 \hat{s}} &> \frac{2\alpha T\check{s}\varepsilon}{T(T-1)\left|X\right|^2 \hat{s}} > \frac{K}{T}\end{aligned} \tag{9}$$

for $K = \frac{2\check{s}\alpha\varepsilon}{\left|X\right|^2\hat{s}}$, which is independent of T and of t.

Before proceeding we show that a similar bound holds in the case in which memory decays, i.e., $\beta < 1$. In this case the total weight of all case-based hypotheses is

$$\begin{aligned}\phi_T\left(\mathcal{CB}_T\right) &= c_T \sum_{i<t<T} \beta^{t-i} \sum_{x,z\in X} s(x,z)\\ &\leq c_T\hat{s} \sum_{i<t<T} \beta^{t-i}\\ &= c_T\hat{s}\frac{\beta}{1-\beta} \sum_{t=0}^{T-1} \left(1-\beta^t\right)\\ &\leq c_T\hat{s}\frac{\beta}{1-\beta} T\end{aligned}$$

and $\phi_T\left(\mathcal{CB}_T\right) > \varepsilon$ implies that

$$c_T\hat{s}\frac{\beta}{1-\beta} T > \varepsilon$$

or

$$c_T > \frac{\varepsilon\,(1-\beta)}{\hat{s}\beta T}$$

To compute the total weight of all case-based hypotheses consistent with h_t^*, we observe that each period $i < t$ contributes one such hypothesis, with weight $c_T\beta^{t-1}s(x,z)$ for some $x, z \in X$ (hence, $s(x,z) \geq \check{s}$). Thus,

$$\begin{aligned}
\phi_T\left(\mathcal{CB}_T \cap R(h_t^*)\right) &\geq c_T \sum_{i=0}^{t-1} \beta^{t-1}\check{s} \\
&= c_T\check{s}\frac{\beta}{1-\beta}\left(1-\beta^t\right) \\
&> \frac{\varepsilon\,(1-\beta)}{\hat{s}\beta T}\check{s}\frac{\beta}{1-\beta}\left(1-\beta^t\right) \\
&= \frac{\varepsilon\check{s}\left(1-\beta^t\right)}{\hat{s}T} \geq \frac{\varepsilon\check{s}\,(1-\beta)}{\hat{s}T}
\end{aligned}$$

and thus the bound

$$\phi_T\left(\mathcal{CB}_T \cap R(h_t^*)\right) > \frac{K}{T}$$

holds for $K = \frac{\varepsilon\check{s}(1-\beta)}{\hat{s}}$.

Since (8) yields

$$\phi_T\left(\mathcal{B}_T \cap R(h_t^*)\right) \leq \frac{P(T)}{\left(d^\alpha\right)^T}$$

and we just showed that

$$\phi_T\left(\mathcal{CB}_T \cap R(h_t^*)\right) > \frac{K}{T}$$

where K is independent of T and of t, we can conclude that

$$\frac{\phi_T\left(\mathcal{B}_T \cap R(h_t^*)\right)}{\phi_T\left(\mathcal{CB}_T \cap R(h_t^*)\right)} < \frac{P(T)T}{K\left(d^\alpha\right)^T}$$

As the numerator is a polynomial in T, and the denominator is an exponential function with base $d^\alpha > 1$, for a large enough T_0 this ratio will be below the pre-specified δ. ■

References

[1] Hirotugu Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6(2):7127–132, 1954.

[2] Enriqueta Aragones, Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler. Fact-free learning. *Ameican Economic Review*, 95(5):1355–1368, 2005.

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

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

[5] W. R. F. Browning. *Parables: A Dictionary of the Bible*. Oxford University Press, Oxford, 1997.

[6] Rudolf Carnap. Uber die aufgabe der physik und die andwednung des grundsatze der einfachstheit. *Kant-Studien*, 28:90–107, 1923.

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

[8] Gustave Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5: (Grenoble):131–295, 1953–54.

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

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

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

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

[13] Evelyn Fix and J. L. Hodges. Discriminatory analysis. Nonparametric discriminiation: Small sample performance. Report a193008, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.

[14] Peter Gärdenfors. Induction, conceptual spaces and AI. *Philosophy of Science*, 57(1):78–95, 1990.

[15] Gerd Gigerenzer, Peter M. Todd, and The ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford, 1999.

[16] Itzhak Gilboa. *Theory of Decision under Uncertainty*. Cambridge University Press, Cambridge, 2009.

[17] Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler. Rationality of belief. Or why Bayesianism is neither necessary nor sufficeint for rationality. Cowles Foundation Discussion Paper 1484, Yale University, 2004.

[18] Itzhak Gilboa and Larry Samuelson. Subjectivity in inductive inference. Cowles Foundation Discussion Paper 1725, Tel Aviv University and Yale University, 12009.

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

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

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

[22] Nelson Goodman. *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, Massachusetts, 1954.

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

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

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

[26] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

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

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

[29] Drew McDermott and John Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13(1), 1980.

[30] Nils J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

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

[32] Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson and Co., London, 1958, reprinted 1961 (New York: Science Editions), originally Logic der Forschung (1934).

[33] 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.

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

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

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

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

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

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

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

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

[42] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 221(4481):453–458, 1981.

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