Analogies and Theories: The Role of Simplicity and the Emergence of Norms*

Gabrielle Gayer†and Itzhak Gilboa‡

September 2012

Abstract

We consider the dynamics of reasoning by general rules (theories) and by specific cases (analogies). When an agent faces an exogenous process, we show that, under mild conditions, if reality happens to be simple, the agent will converge to adopt a theory and discard analogical thinking. If, however, reality is complex, the agent may rely on analogies more than on theories. By contrast, when the agent is a player in a large population coordination game, and the process is generated by all players' predictions, convergence to a theory is much more likely. This may explain how a large population of players selects an equilibrium in such a game, and how social norms emerge. Mixed cases, involving noisy endogenous processes are likely to give rise to complex dynamics of reasoning, switching between theories and analogies.

1 Introduction

Consider a set of agents who attempt to predict the process that governs the environment they live in. They might be facing a process that is *exogenous*, that it, independent of the agents' predictions, or *endogenous*, that

*Gilboa gratefully acknowledges ISF Grant 396/10 and ERC Grant 269754.
†Bar-Ilan University. gabi.gayer@gmail.com
‡HEC, Paris, and Tel-Aviv University, tzachigilboa@gmail.com.

is, fully determined by these predictions. For example, natural processes such as the weather, earthquakes, or hurricanes are exogenous. On the other hand, social processes such as the adoption of a social norm, are largely endogenous, as they are fundamentally determined by the agents' predictions thereof. Many other processes are combination of exogenous and endogenous processes. These include, for example, prices in real estate, commodities, and financial markets, which respond both to exogenous news and to speculative trade.

How do agents reason about such processes? Do they think about exogenous and endogenous processes in the same way? This paper attempts to address these questions and others in a formal way. We consider a dynamic model in which, at each period t, an agent tries to predict the value of a variable y_t, based on a set of observable variables, x_t, as well as the history of both x and y (that is, $(x_i, y_i)_{i<t}$). One common mode of reasoning is regression analysis, whereby y_t might be regressed on x_t, on its own past values, $(y_i)_{i<t}$, or some combination of these. This process belongs to a general category known as *rule-based* learning that involves a selection of theories based on observations. In philosophy, this mode of reasoning is referred to as (case-to-rule) induction, and it is based on the belief that a rule that has been valid in the past will remain valid in the future. Hume (1748) famously pointed out that this belief requires justification, thereby stating the *problem of induction*. Wittgenstein (1922) suggested that the process of induction consists in finding the simplest theory that conforms to the observations, while Goodman (1955) claimed that the notion of simplicity is language-dependent.[1] The basic mechanism of using unrefuted theories for prediction has remained a fundamental method of inference in science, statistics, and everyday life.

Another, perhaps simpler mode of reasoning involves analogical thinking.

[1] Solomonoff (1964) showed that, in an appropriate model, the dependence of simplicity judgments on language can be bounded.

In its simplest manifestation, when the variable x_t is ignored, y_t is predicted to be the most frequently encountered value in the past.[2] If, however, different periods $i < t$ are characterized by different values of x_i, one may wish to rely more heavily on more similar periods,[3] as captured by the statistical techniques of kernel estimation (Akaike, 1954, Parzen, 1962, and see also Silverman, 1986). In artificial intelligence, this mode of reasoning has been referred to as *case-based* (see Schank, 1986, Riesbeck and Schank, 1989), and it has been axiomatized in Gilboa and Schmeidler (2001, 2003). Slade (1991) and Kolodner (1992) pointed out some advantages of case-based systems over rule-based systems.[4]

It appears that both case-based reasoning and rule-based reasoning are common in everyday life, as well as in formal statistical analysis. In the artificial intelligence literature there are attempts to combine the two modes of reasoning in order to exploit their respective advantages (see for example Rissland and Skalak, 1989, and Domingosu, 1996). However, we are unaware of theoretical work that analyzes such combinations, especially as models of human reasoning, dealing with questions such as, when do agents tend to use analogies, and when – theories? Do they converge to one such mode of reasoning in the long run, and if so, which? Or, under which conditions will case-based reasoning be asymptotically dominant, and under which conditions will long run behavior to be governed by rule-based reasoning? Specifically, are there differences between exogenous and endogenous processes in this respect?

We start with an adaptation of the model of Gilboa, Samuelson, and Schmeidler (GSS, 2010), which provides a unified framework for case-based, rule-based, and Bayesian reasoning by assigning a weight to each mode of

[2] Or, in a continuous model, the average value of past observations.

[3] As suggested by Hume (1748, Section IV), "From causes which appear similar we expect similar effects."

[4] Their definition of rule-based systems is, however, different from the definition we use in this paper.

reasoning. The focus of GSS (2010) is the robustness of Bayesian reasoning, versus that of case-based and rule-based reasoning, the main distinction being that the latter two modes of reasoning allow for unquantified uncertainties that are prohibited in Bayesian reasoning. That paper also considers examples of dynamics alternating between rule-based and case-based reasoning.

Rules (or theories) in GSS (2010) can have different domains of applicability, allowing them, in particular, to vary in their starting periods. Thus, a rule in that model might be "Starting at time $t = t_0$, y_t will equal y_0". In this paper, by contrast, we limit attention to theories that share their domain of applicability: they all make predictions at each and every period, predicting y_t given x_t (and given history), but not of x_t itself.[5] We consider a countable set of such theories, presumably all theories that are computable, that is, that can be described by a Turing machine or a PASCAL program. These theories are contrasted with case-based reasoning by examining the long-run behavior of the relative weights of these two modes of reasoning.

The analysis turns out to critically depend on the process that generates the variable y_t. When the process is *exogenous*, namely when y_t is completely independent of the agent's reasoning process, we show that, under mild assumptions, rule-based reasoning will prevail if reality happens to be simple, that is, describable by a Turing machine. However, case-based reasoning will be dominant if reality is complex. Since there are many more complex scenarios than there are simple ones, it is safe to say that, for the most part, rule-based reasoning will wither away making case-based reasoning the only viable mode of reasoning. For example, there are very good models for predicting weather conditions for the following day, but not for a year hence.

[5] Theories in our model are not supposed to predict x_t. Hence they are not "Bayesian" by the definition of GSS (2010). For example, a scientist who uses a function f to generate predictions $y_t = f(x_t)$ would be considered rule-based according to our definition. Such a scientist will not be Bayesian, because, prior to observing x_t, she entertains no beliefs regarding y_t.

This is consistent with the supposition that weather conditions constitute a simple process in the short run (when t is measured in days) but not in the long run (when it is measured in years). Moreover, prediction of weather conditions in the long run typically reduces to past averages, which are a form of case-based reasoning.

If the process is perfectly *endogenous*, where the variable y_t is the mode of the prediction of the agents, it turns out that every scenario is possible. However, mild computability assumptions suggest that case-based reasoning cannot be selected asymptotically. By contrast, rule-based reasoning is likely to be dominant in the long run, because the agents' shared prediction agrees with a certain theory that becomes the theory of choice for their predictions. Thus, when we consider equilibrium selection in a game among many agents, it is more likely to find the agents converging to simple rules than in the case where these agents predict, say, the weather. This convergence to rules may explain the emergence of social norms as the selection of equilibria in coordination games.

As mentioned above, there are many economic phenomena involving intermediate cases, where the process y_t is determined partly by agents' predictions, and partly by exogenous factors. Speculative trade is one such example. In these cases, due to the external "noise" factors, no single theory can remain valid in the long run (unless the noise factors diminish over time). Nevertheless, when the noise factors are relatively weak, it may take a very long time for the process to converge, and in the meantime the agents' reasoning will fluctuate between rule-based and case-based reasoning. In particular, the agents' reasoning may select theories that become the equilibrium prediction for a certain period, until they are refuted, and then replaced by new theories, or by periods of case-based reasoning.

The rest of the paper is organized as follows. Section 2 describes the basic framework. It uses the framework of GSS (2010) and defines rule-based and case-based reasoning. Section 3 deals with a purely endogenous process, showing that rule-based reasoning is likely to emerge in simple states of the world, but not in complex ones. Section 4 then deals with a purely endogenous process, showing that rule-based reasoning is more likely to emerge as the asymptotic mode of reasoning than case-based reasoning. Section 5 concludes with comments on some variations of these models.

2 Framework

2.1 The unified model

We adapt the unified model of induction of GSS (2010). An agent makes predictions about the value of a variable y based on some observations x. She has a history of observations of past x and y values to rely on. We make no assumptions about independence or conditional independence of the variables across periods, or any other assumption about the data generating process.

Let the set of periods be $\mathbb{T} \equiv \{0, 1, 2, \ldots, t, \ldots\}$. At each period $t \in \mathbb{T}$ there is 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

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

For a state ω and a period t, let $\omega(t) = (\omega_X(t), \omega_Y(t))$ denote the element of $X \times Y$ appearing in period t given state ω. Let

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

denote the history of characteristics and outcomes in periods 0 through $t-1$, along with the period-t characteristic, given state ω. Let H_t denote all possible histories at period t, i.e., $H_t = \{h_t(\omega) \mid \omega \in \Omega\}$. We let (h_t, y) denote the concatenation of the history h_t with the outcome y.

[6] Clearly, x_t may be a vector of characteristics, or explanatory variables, each of which assumes only finitely many values.

In each period $t \in \mathbb{T}$, the agent observes a history h_t and makes a prediction about the period-t outcome, $\omega_Y(t) \in Y$. A *prediction* is a ranking of subsets in Y given h_t.

Predictions are made with the help of conjectures. A conjecture is an event $A \subset \Omega$. A conjecture 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. However, not every subset of Ω may be considered by the agent. Rather, we assume that the agent only conceives of a *countable* subset $\mathcal{A}$ of 2^Ω, referred to as the set of *conjectures*. We explain below why countability is a natural restriction for our purposes. For the time being, we mention that only countable sets are considered, so that summation over such sets will be well-defined.

GSS (2010) show that the notion of conjectures is general enough to capture Bayesian, rule-based, as well as case-based reasoning. Specifically, they assume that the agent has a *model*, which is a function $\phi : \mathcal{A} \to \mathbb{R}_+$, where $\phi(A)$ is interpreted as the weight attached to conjecture A for the purpose of prediction. For a subset of conjectures $\mathcal{D} \subset \mathcal{A}$, ϕ is defined additively, that is,

$$\phi(\mathcal{D}) = \sum_{A \in \mathcal{D}} \phi(A).$$

It sacrifices no generality to assume that $\phi(\mathcal{A}) = 1$.[7]

For a history $h_t \in H_t$, define

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

Thus, $[h_t]$ is the event consisting of all states that are compatible with the history h_t. Similarly, for $h_t \in H_t$ and a subset of outcomes $Y' \subset Y$, we define the event

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

[7] In GSS (2010), the set of conjectures is uncountable, and ϕ is defined as a measure over subsets of conjectures, that is, subsets of subsets of states of the world. This complication is obviated thanks to the countability assumption.

consisting of all states that are compatible with the history h_t and with the next outcome being in the set Y'.

The agent learns by ruling out conjectures that have been refuted by evidence. Specifically, given a history $h_t \in H_t$, a conjecture A that is disjoint from $[h_t]$ should not be taken into consideration in future predictions. Fixing a subset of conjectures $\mathcal{D} \subset \mathcal{A}$, a history $h_t \in H_t$ and a subset of outcomes $Y' \subset Y$, consider the set of conjectures in $\mathcal{D}$ that have not been refuted by h_t and that predict that the outcome will be in Y':

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

Observe that the conjectures in $\mathcal{D}(h_t, Y')$ are various events, many pairs of which may not be disjoint. This is important to bear in mind in the following definitions, where we sum over the weights assigned to different conjectures.

Given a model $\phi : \mathcal{A} \to \mathbb{R}_+$, the weight assigned to Y' by the unrefuted conjectures in $\mathcal{D}$ is

$$\phi(\mathcal{D}(h_t, Y')).$$

The total weight assigned to a subset $Y' \subset Y$ by all unrefuted conjectures is thus given by

$$\phi(\mathcal{A}(h_t, Y')).$$

The agent's prediction is a ranking of the subsets of Y, with Y' considered more likely than Y'' iff

$$\phi(\mathcal{A}(h_t, Y')) > \phi(\mathcal{A}(h_t, Y'')).$$

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

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

Observe that $\mathcal{D}(h_t)$ is the set of conjectures in $\mathcal{D}$ that have not been refuted and that could lend their weight to *some* nontautological prediction after history h_t (and hence $\mathcal{D}(h_t) \subset \mathcal{D}(h_t, Y)$.)

2.2 Rule-based reasoning: theories

The notion of a rule is rather general. There are *association rules*, which, conditional on the value of x_t, restrict the possible values of y_t. For example, the rule "if the Democratic candidate wins the election, taxes will rise" says something about the rate of taxation, y_t if the president is a Democrat (i.e., if x_t assumes a certain value). Such a rule does not restrict prediction in case its antecedent does not hold. By contrast, there are *functional rules*, which predict that y_t be equal to $f(x_t)$ for a certain function f. Other rules may be time-dependent, and allow y_t to be a function of x_t as well as of t itself. Further, rules may differ in their domain. In particular, GSS (2010) provide an example of rule-based reasoning in which the rules predict a certain constant y value beginning with a given period t, and making no predictions prior to that t.

In this paper we restrict attention to rules that can be viewed as general theories. Such theories are constrained to make a specific prediction (i.e., a single y_t) at each and every t, and for any possible value of x_t. Moreover, we will allow such functions to depend on the entire history h_t, and thus on previous values (x_i, y_i) for $i < t$. However, we make one important assumption: all the functions we consider are computable by Turing machines. That is, we consider only those theories $f : \cup_{t\geq 0} H_t \to Y$ for which there exists a Turing machine (or, equivalently, a PASCAL program), which, for every t and every h_t, halts in finite time. This appears to be a minimal requirement because a theory that does not halt will fail to compute the value $y_t = f(h_t) \in Y$ for every t. It is well known that there is only a countable number of such theories. We denote the set of *theories* by $\mathcal{R} = \{f_1, f_2, ...\}$.

Observe that the definition assumes that a theory $f_j \in \mathcal{R}$ computes a prediction for every history h_t, including histories that are inconsistent with f_j itself. This is reminiscent of the definition of a strategy in extensive form games. Alternatively, one may restrict the domain of a theory f only to the histories that do not contradict it.[8]

One may wish to enrich the model by introducing Turing machines (or computer programs) explicitly. In this case, each theory f will be represented by infinitely many machines, which are observationally equivalent. The agent will not be able, in general, to tell which machines are equivalent, but equivalent machines will be refuted at the same histories, and thus their impact on predictions will be the same as that of the function f they represent.

To avoid problems related with undecidability,[9] one may restrict attention to a subset of theories that can be proven to always halt. As long as the subset considered is sufficiently rich to be able to describe any finite history, our results will hold.

If there are no x values to be observed (that is, $|X| = 1$), then for every $f_j \in \mathcal{R}$, there exists a unique state of the world compatible with it. In this case, a model ϕ that puts positive weight only on theories in $\mathcal{R}$ can also be viewed as a Bayesian model (as defined in GSS, 2010), namely as a model assigning probabilities to single states.[10] However, in the more general case, a theory $f_j \in \mathcal{R}$ is compatible with a non-singleton conjecture, because such a theory, as opposed to a Bayesian conjecture, need not predict the values of the x_t's.

For a model ϕ and a theory $f_j \in \mathcal{R}$, we will use $\phi(f_j)$ to denote the weight assigned by ϕ to the conjecture consisting of all the states that do not contradict f_j, that is, $\phi(f_j) = \phi([f_j])$ where

$$[f_j] = \{\omega \in \Omega \mid \omega_Y(t) = f_j(h_t) \quad \forall t\}.$$

[8] Such a restriction would not make a major difference, because the definition of a theory at histories incompatible with it will be immaterial for our purposes. Clearly, a computable theory that is defined on the restricted domain can be extended to a computable theory on the entire domain, say, by predicting a constant y for all histories that are incompatible with the theory.

[9] This is known as the "Halting Problem" by which there is no general method to determine whether a program will halt in finite time.

[10] The resulting Bayesian prior, however, is restricted to have a countable support consisting of computable states.

A model ϕ_R is (a priori) *purely rule-based* if $\phi_R(\mathcal{R}) = 1$, equivalently, $\phi_R(\mathcal{A}\backslash\mathcal{R}) = 0$ or $\sum_j \phi_R(f_j) = 1$. Such a model can also be viewed as a probability distribution over $\mathcal{R}$.

2.3 Case-based reasoning: analogies

Case-based conjectures are defined as in GSS (2010): for every $i < t$, $x, z \in X$, let

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

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

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

$$\mathcal{CB} = \{A_{i,t,x,z} \mid i < t, x, z \in X\} \subset \mathcal{A}. \quad (1)$$

A model ϕ_{CB} is a priori *purely case-based* if all weight is put on the case-based conjectures. Our main interest will be, however, in the evolution of the relative weight of case-based and rule-based reasoning over time, considering the ratio $\frac{\phi(\mathcal{R}(h_t(\omega)))}{\phi(\mathcal{CB}(h_t(\omega)))}$ as a function of t at different states ω.

For example, the agent might have a similarity function over the characteristics,

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

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

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

where **1** is the indicator function of the subscripted event. Hence, the agent may be described as if she considered past cases in the history h_t, chose all those that resulted in some period i with the outcome y, and considered the aggregate similarity of the respective characteristic $\omega_X(i)$ to the current characteristic $\omega_X(t)$. The resulting sums $S(h_t, y)$ can then be used to rank the possible outcomes y. If $\beta = 1$ and in addition the similarity function is constant, the resulting number $S(h_t, y)$ is proportional to the relative empirical frequency of y's in the history h_t.

As noted by GSS (2010), for every similarity function s and decay factor β one may define a model $\phi_{s,\beta}$ by setting $\phi_{s,\beta}(A_{i,t,x,z})$, for each t, to be proportional to $\beta^{(t-i)}s(x,z)$, and $\phi_{s,\beta}(\mathcal{A}\backslash\mathcal{CB}) = 0$. In this case, for every history h_t and every $y \in Y$, $\phi_{s,\beta}(\mathcal{A}(h_t, \{y\}))$ is proportional to $S(h_t, y)$. Such a model $\phi_{s,\beta}$ will be equivalent to case-based prediction according to the function S.

2.4 Open-Mindedness

We restrict our agent to a specific type of rule-based reasoning and a similarly specific type of case-based reasoning. Formally, we assume that the set of conjectures is $\mathcal{A} = \mathcal{R} \cup \mathcal{CB}$. Within this constraint, we wish to guarantee that the agent is open-minded. Thus, we will henceforth assume that the agent assigns a positive weight $\phi(A) > 0$ to each conjecture in $\mathcal{A} = \mathcal{R} \cup \mathcal{CB}$. We denote this set of *open-minded* models by Φ_+.

3 Exogenous Process

3.1 Simplicity Result

For each theory $f_j \in \mathcal{R}$, recall that $[f_j]$ is the event in which f_j is never refuted. All states $\omega \in [f_j]$ are simple in a certain sense: the computation of y_t given h_t can be done in a finite time, employing a program that is independent of t. Observe that the pattern of x_t's in ω may be rather complicated, and, in particular, it can be a pattern that cannot be computed by any Turing machine. However, since the agent's task is to predict y_t *given* h_t, we ignore this complexity. We therefore define the *set of simple states* to be

$$\mathbb{S} = \bigcup_{r \geq 1} [f_r]$$

(or $\mathbb{S} = \bigcup_{f \in \mathcal{R}} [f]$).

We can now state

Proposition 1 *For every $\phi \in \Phi_+$ and every $\omega \in \mathbb{S}$,*

$$\frac{\phi\left(\mathcal{CB}(h_t\left(\omega\right))\right)}{\phi\left(\mathcal{R}(h_t\left(\omega\right))\right)} \rightarrow 0$$

as $t \rightarrow \infty$.

That is, in all simple states, the agent will converge to reason by theories and will gradually discard case-based reasoning.

The logic of this proposition is straightforward: if we consider a simple state ω, where a certain simple theory f_r holds, the initial weight assigned to this theory will serve as a lower bound on $\phi\left(\mathcal{R}(h_t\left(\omega\right))\right)$ for all t, because the theory will never be refuted at ω. By contrast, the total weight of the set of all case-based conjectures that are relevant for prediction at time t converges to zero because it is an element in a convergent series. Intuitively, because at ω the theory f_r is correct, it retains its original weight of credence. By contrast, case-based conjectures concern only pairs of periods, $i < t$, and thus, for each new value of t, a new set of case-based conjectures is being considered. It is inevitable that the total weight of this set (which is disjoint from sets considered in previous periods) converge to zero.

3.2 The Fragility of Rule-Based Reasoning

Because there are only countably many simple states of the world, it is intuitive that "most" states are not simple. What happens in these states?

Clearly, in such a state no single theory can be unrefuted forever. But it is still possible that different theories succeed each other in the agent's mind, so that at each period her reasoning is mostly rule-based.

While this is possible in some non-simple states, and even in many of them, there is still a well-defined sense in which it can only happen in a small minority of states. For example, assume that $|X| = 1$ and that $Y = \{0,1\}$, so that the state space is $\{0,1\}^{\mathbb{N}}$. Let λ be the measure defined by assigning weight 2^{-t} to any event defined by $(y_0, ..., y_{t-1})$ (for any sequence of t values for the t variables). In this case, one can show that the set of states where non-negligible weight is put on rule-based reasoning is small, as measured by λ. More generally, we will prove this result for any finite sets X and Y, and for a large class of measures λ.

Endow the state space Ω with the σ-algebra Σ defined by the variables $(x_t, y_t)_{t \geq 0}$. A probability measure λ on Σ is a *non-trivial conditionally iid measure* if, for every $x \in X$ there exists $\lambda_x \in \Delta(Y)$ such that (i) for every $h_t = ((x_0, y_0), \ldots, (x_{t-1}, y_{t-1}), x_t)$, the conditional distribution of Y given h_t according to λ is λ_x; and (ii) λ_x is non-degenerate for every $x \in X$. The measure λ is assumed neither to govern the actual process, nor to capture the reasoner's beliefs. It is merely a way to quantify states of the world, and capture the intuition that certain events are small relative to others.

Proposition 2 *Let there be given $\phi \in \Phi_+$ and let λ be a non-trivial conditionally iid measure. For every $\varepsilon > 0$ there exists T_0 such that*

$$\lambda\left(\left\{\omega \,\middle|\, \phi\left(\mathcal{R}(h_t(\omega))\right) \leq \delta^{t/2} \quad \forall t \geq T_0\right\}\right) > 1 - \varepsilon.$$

This result states that, apart from a λ-negligible event, the weight of the rule-based conjectures decreases at a semi-exponential rate. Clearly, this cannot be the case in the simple state, where the weight of the rule-based conjectures remains bounded away from zero. But there are only countably many simple states, and they are therefore of λ-measure zero. Thus, there can be many non-simple states at which the weight of the rule-based conjectures does not decay very fast, but the total (λ-)weight of all these states, simple or non-simple, is negligible.

Does the fast decay of the weight of the rule-based conjectures mean that the reasoner will tend to use more case-based conjectures? The answer depends on the rate at which the weight of the case-based conjectures tends to zero. Thus, we are led to ask, how are the weights to be spread over the case-based conjectures?

One may argue that it is intuitive for the total weight of the case-based conjectures at a given time to be independent of t. However, the set of case-based conjectures that are relevant at t is disjoint from the corresponding set for $t' \neq t$. It is therefore a mathematical necessity that the weight assigned to all case-based conjectures relevant at period t converge to zero (as in the proof of Proposition 1). However, there is no reason for this total weight to converge to zero too fast. We therefore assume that the weight of all case-based conjectures, across all periods, is split among them so that the case-based conjectures relevant for prediction at each history h_t command a positive weight that does not diminish too fast as a function of t. This will be the case if, for instance, the total weight is split proportionately to a strictly positive similarity matrix $S : X^2 \to \mathbb{R}$.

Formally, define $\Phi_+^p \subset \Phi_+$ to be the set of models ϕ for which there exist $\gamma < -1$ and $c > 0$, such that, for every t, and every $x, z \in X$,

$$\sum_{i<t} \phi\left(A_{i,t,x,z}\right) \geq ct^{\gamma}$$

Under these assumptions, the opposite of that of Proposition 1 holds almost everywhere:

Proposition 3 *Let there be given a model $\phi \in \Phi_+^p$ and a non-trivial conditionally iid measure λ. Then, λ-almost everywhere,*

$$\frac{\phi\left(\mathcal{R}(h_t\left(\omega\right))\right)}{\phi\left(\mathcal{CB}(h_t\left(\omega\right))\right)} \to 0$$

as $t \to \infty$.

4 Endogenous Process

In this section we consider a process that is governed by the reasoning of a set of agents. For example, consider the behavior of agents involved in a coordination game, where each agent tries to predict the social norm that will govern the behavior of others, and to match that norm in her choice of strategy.

In this section we analyze the case in which all agents share the same weight function $\phi \in \Phi$. Beyond serving as an important benchmark, this extreme case attempts to capture the intuition that, while people vary in their a priori judgment of theories, these judgments are correlated. Specifically, people tend to prefer simpler theories to more complex ones, and similarity judgments are correlated across people. For example, people might disagree whether the pattern 011111... is simpler than the pattern 010101..., but practically everyone would agree that 000000... is simpler than 011001... (where readers might wonder how the last sequence is meant to be continued). For such a function ϕ define

$$\Omega_\phi = \left\{ \omega \in \Omega \;\middle|\; \omega_Y(t) \in \arg\max_{y \in Y} \phi(\mathcal{A}(h_t, \{y\})) \quad \forall t \geq 0 \right\}.$$

Thus, it is assumed that the agents' predictions determine the actual outcome. As in the case of the exogenous process, the agents are not assumed to predict the values of x_t, nor to affect them.

We first note that every state of the world may unfold in an endogenous process:

Proposition 4 *For every $\omega \in \Omega$, there exists $\phi \in \Phi_+^p$ such that $\omega \in \Omega_\phi$.*

The proof of Proposition 4 is constructive: given a state $\omega \in \Omega$ the proof describes an algorithm that generates $\phi \in \Phi_+^p$ such that $\omega \in \Omega_\phi$. However, if the state ω itself is not computable, the resulting ϕ may also not be computable. Hence the message of the proposition should be qualified:

while every state may materialize as the prediction of the agents, if we make the plausible assumption that the agents can only make predictions using a computable ϕ, not all states are necessarily possible.

We therefore restrict attention to models ϕ that are computable functions assuming rational values. That is, consider only the set of functions $\Phi_+^{cp} \subset \Phi_+^p$ such that, for each $\phi \in \Phi_+^{cp}$, there is a machine that computes $\phi(A) \in \mathbb{Q}$ for each conjecture A. This assumption allows us to imagine the agent as actually trying to compute $\phi(A(h_t, \{y\}))$ for each $y \in Y$, and choosing an ε-maximizer of $\phi(A(h_t, \{\cdot\}))$ as her prediction.[11] Clearly, this interpretation is not the only one, and one may think of the computation of $\phi(\mathcal{A}(h_t, Y'))$ as a model that an outside observer uses in order to provide a description of the agent's predictions.

Our interest is in the dynamics of reasoning of the agents along states in Ω_ϕ for $\phi \in \Phi_+^{cp}$. To this end, we introduce the following definitions. *Rule-based reasoning is dominant* at state $\omega \in \Omega_\phi$ at period t if

$$\begin{array}{rrcl} (i) & \phi\left(\mathcal{R}(h_t\left(\omega\right))\right) & > & \phi\left(\mathcal{CB}(h_t\left(\omega\right))\right) \\ & & & \textit{and} \\ (ii) & \omega_Y\left(t\right) & \in & \arg\max_{y \in Y} \phi(\mathcal{R}(h_t, \{y\})). \end{array}$$

Thus, rule-based reasoning is dominant if there is more weight put on rule-based reasoning than on case-based reasoning, and if the prediction of the rule based reasoning is indeed the prediction that the agents make (and that defines the next observation y_t). Similarly, we say that *case-based reasoning*

[11] Observe that the computation of $\phi(A(h_t, \{y\}))$ involves infinite summations. Hence the agent cannot simply compute $\phi(A(h_t, \{y\}))$ for each y with perfect precision. However, the agent can be imagined to simultaneously approximate these values and halt the computation if the difference between the values is larger than the residual weight, or if the residual weight is below a certain threshold. This would result in a computable procedure that approximates the maximization ϕ in the sense that it provides an ε–maximization of ϕ

is dominant at state $\omega \in \Omega_\phi$ at period t if

$$\begin{array}{rcl} (i) \quad \phi\left(\mathcal{R}(h_t(\omega))\right) & < & \phi\left(\mathcal{CB}(h_t(\omega))\right) \\ & & \textit{and} \\ (ii) \quad \omega_Y(t) & \in & \arg\max_{y \in Y} \phi(\mathcal{CB}(h_t, \{y\})). \end{array}$$

Observe that, at $\omega \in \Omega_\phi$ at period t we may have neither mode of reasoning dominating either if they happen to be equally weighty, that is, if, $\phi\left(\mathcal{R}(h_t(\omega))\right) = \phi\left(\mathcal{CB}(h_t(\omega))\right)$, or if the weightier mode of reasoning does not correctly predict the outcome. This may happen, for instance, if the conjectures in the dominant mode of reasoning split the weight between the different predictions, so as to make the other mode of reasoning pivotal.

For $\phi \in \Phi_+^{cp}$ we are interested in the long-run existence of a dominant mode of reasoning. Define $\Omega_{RB\phi}$ to be the set of states $\omega \in \Omega_\phi$ such that, for some T, rule-based reasoning is dominant at state $\omega \in \Omega_\phi$ at all $t \geq T$. Define $\Omega_{CB\phi}$ accordingly to be the states at which case-based reasoning dominates from some period on.

Proposition 5 *For every $\phi \in \Phi_+^{cp}$ we have $\mathbb{S} \subset \Omega_{RB\phi}$.*

Thus, for every weight function that satisfies our assumptions, the set of states in which rule-based reasoning is eventually dominant contains all the simple states. One might wonder whether in complex states case-based reasoning might be dominant in the long run. The negative answer is given by

Proposition 6 *For every $\phi \in \Phi_+^{cp}$ we have $\Omega_{CB\phi} = \varnothing$.*

The reasoning behind Proposition 6 is very simple: if ω were a state that is, in the long run, governed by case-based reasoning, then, because ϕ is computable, there exists a theory that simulates the case-based reasoning defined by ϕ. For example, if all agents simply use the modal y for prediction, there exists a simple algorithm that describes their prediction, and therefore the resulting state ω. Since an open-minded ϕ must have assigned this algorithm a positive weight a priori, the theory described by this algorithm will eventually prevail as the correct theory used for prediction.

By the same logic, one might be tempted to suggest that, for any computable ϕ, $\Omega_{RB\phi} = \Omega_{\phi}$, that is, that the process ends up in a rule-based state. This conclusion would not be warranted for two reasons: (i) the condition $\omega_Y(t) \in \arg\max_{y \in Y} \phi(\mathcal{A}(h_t, \{y\}))$ does not require that $\omega_Y(t)$ be a single maximizer of $\phi(\mathcal{A}(h_t, \{y\}))$; in case of ties, ω may involve a pattern of choices of y that is not computable; and, moreover, (ii) as mentioned above, computability of ϕ does not imply that $\phi(\mathcal{A}(h_t, \{y\}))$ is computable, as the latter involves an infinite summation. (This cannot happen when one restricts attention to case-based conjectures, but it will necessarily be the case when rule-based conjectures are concerned.)

5 Variants

5.1 Hybrid models

Consider the case of trade in financial markets. Financial assets are affected by various economic variables that are exogenous to the market, ranging from weather conditions to technological innovation, from demand shocks to political revolutions. At the same time, financial assets are worth what the market "thinks" they are worth. In other words, such markets have a strong endogenous factor as well. It seems natural to assume that such processes (y_t) are governed partly by the predictions $(\hat{y}_t)$ as in Section 4 and partly by random shocks as in Section 3. For instance, assume that $\alpha(h_t)$ is the probability that agents' reasoning determines y_t, and with the complement probability y_t is determined by a random shock. That is,

$$y_t = \begin{cases} \hat{y}_t & \alpha(h_t) \\ \tilde{y}_t & 1 - \alpha(h_t) \end{cases}$$

where $\hat{y}_t \in \arg\max_{y \in Y} \phi(\mathcal{A}(h_t, \{y\}))$ and $\tilde{y}_t$ is uniformly distributed over Y. Thus, if $\alpha(h_t) \equiv 1$ we consider a model as in Section 4, which is likely to converge to a single dominant theory, and when $\alpha(h_t) \equiv 0$ we consider a model as in Section 3, coupled with a non-degenerate iid measure that guarantees asymptotic case-based reasoning. Obviously, the interesting case is where $\alpha(h_t) \in (0, 1)$ (for most if not all histories h_t).

If $\alpha(h_t)$ is independent of history, so that $\alpha(h_t) \equiv \alpha \in (0, 1)$, no theory can be dominant asymptotically. Indeed, every theory that correctly predicts $\hat{y}_t$ has a fixed positive probability $(1 - \alpha)$ of being refuted at each period, and will thus be refuted at some point with probability 1. Moreover, when t is large, we know that with very high probability the number of "noise" periods is approximately $(1 - \alpha)t$. Over these periods we are likely to observe a complex pattern of y_t's, and thus a result similar to Proposition 3 holds: the total weight of rule-based conjectures decreases, on average, exponentially fast in the number of noise periods. Because the number of noise periods increases linearly in t (as it is roughly $(1 - \alpha)t$), this weight is also an exponentially decreasing function of t and thus it decays faster than do the case-based conjectures. Thus, case-based reasoning will be asymptotically dominant in "most" states of the world even if $\alpha(h_t) \equiv \alpha$ is very close to 1.

However, the probability of noise in an endogenous process is likely to be endogenous as well. For example, consider the choice of driving on the right or on the left in a large population. When agents are not quite sure which equilibrium is being played, it is easier for a random shock to switch equilibria. But when all the agents are rather certain that everyone is going to drive, say, on the right, it is highly unlikely that at least half of them would behave differently from what they would find optimal based on their predictions. Thus, it stands to reason that $\alpha(h_t)$ depends on h_t, and, moreover, that it converges to 1 as t grows, if a simple theory fits the data h_t. Such convergence would allow the process to be asymptotically dominated by rule-based conjectures with positive probability.

5.2 Heterogenous beliefs

The analysis in Section 4 assumes that all agents share the function ϕ, which is the natural counterpart of the common prior assumption in economics. Clearly, this assumption is not entirely realistic; people vary in their similarity judgments, in their prior beliefs in theories, as well as in their tendency to reason by theories vs. by analogies. Hence one may consider an endogenous process in which the population is distributed among different credence functions ϕ.

Importantly, the distinction between computable and incomputable states is an objective one. Agents may vary in the language they use to describe theories, and, correspondingly, in their judgment of simplicity. However, any two languages that are equivalent to the computational model of a Turing machine can be translated to each other. Thus, if the process follows a simple (computable) path, all agents will notice this regularity. Different agents may discard case-based reasoning in favor of the unrefuted theory at different times, but (under the assumption of open-mindedness) all of them will eventually realize that this unrefuted theory is indeed "correct". Interesting dynamics might emerge if the agents who are slow to switch to prediction by the correct theory are sufficiently numerous to refute that theory, thereby changing the reasoning of those agents who were the first to adopt the theory.

6 Appendix: Proofs

6.1 Proof of Proposition 1

Assume that $\omega \in [f_r]$ for some r. In this case the denominator is bounded from below by the weight assigned to the correct theory f_r. In fact,

$$\mathcal{R}(h_t(\omega)) \searrow \phi(f_r) > 0$$

as $t \to \infty$.

By contrast, $\mathcal{CB}(h_t(\omega))$ includes the ϕ-weight only of those case-based conjectures that are relevant at t, that is

$$\phi(\mathcal{CB}(h_t(\omega))) = \sum_{\{(i,x,z)|i<t,\omega_X(i)=x,\omega_X(t)=z\}} \phi(A_{i,t,x,z})$$

Clearly,

$$\phi(\mathcal{CB}(h_t(\omega))) \leq \sum_{\{(i,x,z)|i<t\}} \phi(A_{i,t,x,z})$$

Defining

$$\alpha_t = \sum_{\{(i,x,z)|i<t\}} \phi(A_{i,t,x,z})$$

and observing that

$$\sum_t \alpha_t = \phi(\mathcal{CB}) < 1$$

we must have

$$\alpha_t \to 0$$

as $t \to \infty$. Hence $\phi(\mathcal{CB}(h_t(\omega)))$ also converges to zero as $t \to \infty$, and this completes the proof.

6.2 Proof of Proposition 2

Let there be given an open-minded model ϕ. For a period t and a sequence $x_{(t)} = (x_0, ..., x_{t-1}) \in X^t$, consider the state space $\Omega_{x_{(t)}}$ defined by the corresponding $y_{(t)} = (y_0, ..., y_{t-1}) \in Y^t$ and containing $|Y|^t$ states. Thus $\Omega_{x_{(t)}}$ is a replica of Y^t and when no confusion is likely to arise we will refer to elements of $\Omega_{x_{(t)}}$ as $y_{(t)}$.

Let $\delta < 1$ be such that $\lambda_0((x, y)) < \delta$ for every $(x, y) \in X \times Y$. Observe that λ attaches a probability not exceeding δ^t to each element in the space $\Omega_{x_{(t)}}$.

Let W be a random variable defined on $\Omega_{x_{(t)}}$, and measuring the total weight of rule-based conjectures that are compatible with history. That is, for $y_{(t)} = (y_0, ..., y_{t-1})$ choose an arbitrary $x_t \in X$ and define h_t by $h_t = ((x_0, y_0), ..., (x_{t-1}, y_{t-1}), x_t)$. Choose ω such that $h_t(\omega) = h_t$ and define

$$W(y_{(t)}) = \phi(\mathcal{R}(h_t(\omega)))$$

Clearly, such states ω exist. Importantly, $\mathcal{R}(h_t(\omega))$ does not depend on the choice of x_t because theories are not required to predict x_t and therefore no observation of x_t will rule out any theories. It is also obvious that the choice of ω such that $h_t(\omega) = h_t$ does not affect $\mathcal{R}(h_t(\omega))$, because yet-unobserved variables do not rule out theories.

Observe that $\{\mathcal{R}(h_t(\omega))\}_\omega$ defines a partition of $\mathcal{R}$: each theory $f \in \mathcal{R}$ is compatible with precisely one state $y_{(t)} \in \Omega_{x_{(t)}}$. Hence

$$\sum_{y_{(t)} \in \Omega_{x_{(t)}}} \phi(\mathcal{R}(h_t(\omega))) = r < 1$$

and therefore

$$\begin{aligned} E(W) &= \sum_{y_{(t)} \in \Omega_{x_{(t)}}} \lambda_{|x_{(t)}}(y_{(t)}) W(y_{(t)}) \\ &= \sum_{y_{(t)} \in \Omega_{x_{(t)}}} \lambda_{|x_{(t)}}(y_{(t)}) \phi(\mathcal{R}(h_t(\omega))) \\ &< \delta^t r < \delta^t. \end{aligned}$$

Denoting by B_t the event $W > \delta^{t/2}$, and using Markov's inequality, we get

$$\lambda_{|x_{(t)}}(B_t) = \lambda_{|x_{(t)}}\left(W > \delta^{t/2}\right) < \frac{E(W)}{\delta^{t/2}} < \frac{\delta^t}{\delta^{t/2}} = \delta^{t/2}.$$

We will also use B_t to denote the corresponding event in Ω. Since we have shown that $\lambda_{|x_{(t)}}(B_t) = \lambda\left(B_t | x_{(t)}\right) < \delta^{t/2}$ for all $x_{(t)}$, we also have $\lambda(B_t) < \delta^{t/2}$.

Next observe that the bounds on the probabilities of the various B_t events converge. In fact,

$$\sum_{t \geq T} \delta^{t/2} \leq \sum_{t \geq 0} \delta^{t/2} = \frac{1}{1 - \sqrt{\delta}}.$$

This implies that for the given $\varepsilon > 0$ there is a large enough T_0 such that

$$\sum_{t \geq T_0} \delta^{t/2} < \varepsilon$$

and thus, for this T_0,

$$\lambda\left(\cup_{t \geq T_0} B_T\right) < \varepsilon$$

and

$$\lambda\left(\left\{\omega \;\middle|\; \phi\left(\mathcal{R}(h_t(\omega))\right) \leq \delta^{t/2} \quad \forall t \geq T_0\right\}\right) > 1 - \varepsilon.$$

6.3 Proof of Proposition 3

Consider a given $\varepsilon > 0$ and let T_0 be the period provided by Proposition 2. Then, on the corresponding event (whose probability is at least $1 - \varepsilon$)

$$\phi\left(\mathcal{R}(h_t(\omega))\right) \leq \delta^{t/2} \quad \forall t \geq T_0$$

and this, together with the assumption that $\phi \in \Phi_+^p$, that is, $\sum_{i<t} \phi\left(A_{i,t,x,z}\right) \geq ct^{\gamma}$ for $c > 0$ and $\gamma < -1$, implies that

$$\frac{\phi\left(\mathcal{R}(h_t(\omega))\right)}{\phi\left(\mathcal{CB}(h_t(\omega))\right)} < \frac{\delta^{t/2}}{ct^{\gamma}}$$

where the right hand side converges to 0 as t tends to ∞.

Considering a sequence $\varepsilon_n \searrow 0$, one concludes that the convergence to 0 occurs everywhere apart from a set whose λ-measure is zero.

6.4 Proof of Proposition 4

Let there be given $\omega \in \Omega$. For simplicity of notation, we define ϕ without guaranteeing the normalization $\phi(\mathcal{A}) = 1$. It will be obvious from the construction, however, that $0 < \phi(\mathcal{A}) < \infty$ so that ϕ can be normalized.

We first define ϕ on the case-based conjectures. For every $t \geq 1$, let

$$\phi(A_{i,t,x,z}) = \frac{1}{(t+5)^3}.$$

(We use a "lag" of 5 periods to make sure that the rate of decay between any two consecutive periods is not too fast. Specifically, we wish to guarantee that each element in the sequence is at least half of its predecessor.)

Clearly,

$$\phi(\mathcal{CB}) \leq |X|^2 \sum_{t \geq 1} \left[t \frac{1}{(t+5)^3} \right] < \infty.$$

We now turn to define ϕ on $\mathcal{R}$. In the proof we wish to assign weights to subsets of conjectures in $\mathcal{R}$. Note that for every subset $\mathcal{R}' \subset \mathcal{R}$ and every $a > 0$ one may assign a positive weight $\phi(f) > 0$ to each $f \in \mathcal{R}$ such that $\phi(\mathcal{R}') = a$, say by considering an enumeration of $\mathcal{R}'$, $f_1, f_2, ...$ and setting $\phi(f_j) = a/2^j$. In the rest of this proof, we will simply say "assign a weight $a > 0$ to the subset $\mathcal{R}'$", referring to such an assignment.

If $\omega \in \mathbb{S}$, there exists a theory $f \in \mathcal{R}$ such that $\omega \in [f]$. In this case, assign $\phi(f) = 1$ and assign the weight $a = 1/4$ to the set of all the other theories, $\mathcal{R} \backslash \{f\}$. It is easily observed that, at each $t \geq 0$, $\omega_Y(t) \in \arg\max_{y \in Y} \phi(\mathcal{A}(h_t, \{y\}))$ and thus $\omega \in \Omega_\phi$ is established, while $\phi \in \Phi_+^p$ holds.

Next assume that $\omega \notin \mathbb{S}$. Denote, for $t \geq 0$,

$$\mathcal{R}_t = \mathcal{R}(h_t(\omega)).$$

$\mathcal{R}_t$ denotes the set of theories that are unrefuted by history $h_t(\omega)$. Observe that they are all relevant for prediction at period t. Clearly, $\mathcal{R}_0 = \mathcal{R}$, as $h_0(\omega)$ contains only the value of x_0 and no theory makes any prediction about the x's. Moreover, $\mathcal{R}_{t+1} \subset \mathcal{R}_t$, because any theory that agrees with ω for the first $(t+1)$ observations also agrees with it for the first t observations. Finally,

$$\cap_t \mathcal{R}_t = \varnothing$$

because $\omega \notin \mathbb{S}$. We can thus define, for $t > 1$, the set of theories that are proven wrong at period t to be

$$\mathcal{W}_t = \mathcal{R}_{t-1} \backslash \mathcal{R}_t.$$

Observe that

$$\mathcal{R} = \cup_t \mathcal{W}_t$$

and

$$\mathcal{W}_t \cap \mathcal{W}_{t'} = \varnothing$$

whenever $t \neq t'$.

Thus, at period t $\mathcal{R}_t$ consists of all theories that were unrefuted by $h_t(\omega)$, and it is the disjoint union of $\mathcal{R}_{t+1}$, namely the theories that correctly predict $y_t = \omega_Y(t)$ and $\mathcal{W}_{t+1}$, namely the theories that predict different values for y_t, and that will be proven wrong.

If we ignore the case-based conjectures, the prediction made by the theories in $\mathcal{R}_t$ is guaranteed to be the "correct" prediction $\omega_Y(t)$ if

$$\phi(\mathcal{R}_{t+1}) > \phi(\mathcal{W}_{t+1}).$$

(Observe that, as compared to $h_t(\omega)$, $h_{t+1}(\omega)$ specifies two additional pieces of information: the realization of y_t, $\omega_Y(t)$, and the realization of x_{t+1}, $\omega_X(t+1)$. However, theories do not predict the x values, and thus the theories in $\mathcal{R}_{t+1}$ are all those that were in $\mathcal{R}_t$ and that predicted $y_t = \omega_Y(t)$; the observation of x_{t+1} does not refute any additional theories.)

A simple way to construct $\phi \in \Phi_+^p$ is to make sure that the prediction at each period is dominated by the rule-based conjectures, despite the existence of the case-based conjectures. To guarantee that this is the case, we set

$$\phi\left(\mathcal{R}_t\right) = \frac{3}{\left(t+5\right)^2}$$

at each $t \geq 0$.

Observe that, for $t \geq 0$,

$$\begin{aligned}
\phi\left(\mathcal{R}_{t+1}\right) &= \frac{3}{\left(t+6\right)^2} \\
\phi\left(\mathcal{W}_{t+1}\right) &= \phi\left(\mathcal{R}_t\right) - \phi\left(\mathcal{R}_{t+1}\right) \\
&= \frac{3}{\left(t+5\right)^2} - \frac{3}{\left(t+6\right)^2}.
\end{aligned}$$

This dictates the definition of ϕ on $\mathcal{R}$: we start with $\phi\left(\mathcal{R}\right) = \phi\left(\mathcal{R}_0\right) = \frac{3}{5^2}$, and assign the weight $3\left[\left(t+5\right)^{-2} - \left(t+6\right)^{-2}\right]$ to the subset of theories $\mathcal{W}_{t+1}$. Since $\cup_t \mathcal{W}_t = \mathcal{R}$, this defines ϕ on all of $\mathcal{R}$. Clearly, $\phi\left(\mathcal{R}\right)$ is finite.

Next, observe that at each $t \geq 0$, $\omega_Y\left(t\right) \in \arg\max_{y \in Y} \phi(\mathcal{A}(h_t, \{y\}))$. Specifically, at $t = 0$ we only have to compare the rule-based hypotheses. We have

$$\begin{aligned}
\phi\left(\mathcal{R}_1\right) &= \frac{3}{6^2} \\
\phi\left(\mathcal{W}_1\right) &= \frac{3}{5^2} - \frac{3}{6^2}
\end{aligned}$$

so that

$$\phi\left(\mathcal{R}_1\right) - \phi\left(\mathcal{W}_1\right) = 2\frac{3}{6^2} - \frac{3}{5^2} > 0.$$

For each $t \geq 1$, the total weight of the case-based conjectures is

$$t\frac{1}{\left(t+5\right)^3}.$$

We wish to show that the weight of the theories that predict the "correct" continuation $\omega_Y\left(t\right)$, $\mathcal{R}_{t+1}$, is larger than that of the theories that predict other continuations, even when the latter is combined with all case-based conjectures. Indeed,

$$\phi\left(\mathcal{R}_{t+1}\right) - \phi\left(\mathcal{W}_{t+1}\right) = 2\frac{3}{\left(t+6\right)^2} - \frac{3}{\left(t+5\right)^2} > t\frac{1}{\left(t+5\right)^3}.$$

This completes the proof that $\omega_Y(t) \in \arg\max_{y \in Y} \phi(\mathcal{A}(h_t, \{y\}))$ for all t, and it is easily verified that after normalization we obtain $\phi \in \Phi_+^p$ such that $\omega \in \Omega_\phi$. $\square$

6.5 Proof of Proposition 5

Assume that $\omega \in \mathbb{S}$. Then there exists a theory $f \in \mathcal{R}$ such that $\omega \in [f]$. Since $\phi \in \Phi_+$, $\phi(f) > 0$ and this implies that $\phi(\mathcal{R}(h_t(\omega))) > \phi(f) > 0$ for all t. By contrast, $\phi(\mathcal{CB}(h_t(\omega))) \searrow 0$. Similarly, $\phi(\mathcal{R}(h_t(\omega)) \backslash \mathcal{R}(h_{t+1}(\omega))) \searrow 0$ because the sets $\{\mathcal{R}(h_t(\omega)) \backslash \mathcal{R}(h_{t+1}(\omega))\}_t$ are pairwise disjoint (and the sum of their weights is bounded). Hence, from some T onwards, theory f dominates prediction and $\omega \in \Omega_{RB\phi}$. $\square$

6.6 Proof of Proposition 6

Let there be given $\phi \in \Phi_+^{cp}$ and assume that $\omega \in \Omega_{CB\phi}$. This implies that, from some T onwards, $\omega_Y(t)$ can be computed from $h_t(\omega)$ by an algorithm that mimics the summation of ϕ. Because ϕ itself is computable, there exists a theory $f \in \mathcal{R}$ such that $\omega \in [f]$, and it follows that $\omega \in \Omega_{RB\phi}$. Clearly, $\Omega_{RB\phi} \cap \Omega_{CB\phi} = \varnothing$ and it follows that $\Omega_{CB\phi} = \varnothing$. $\square$

7 References

Akaike, H. (1954), "An Approximation to the Density Function", *Annals of the Institute of Statistical Mathematics*, **6**: 127-132.

Domingosu, P. (1996) "Unifying Instance-Based and Rule-Based Induction", Machine Learning, **24**, 141–168

Gilboa, I. and D. Schmeidler (2001), *A Theory of Case-Based Decisions*, Cambridge: Cambridge University Press.

——— (2003) "Inductive Inference: An Axiomatic Approach", *Econometrica*, **71**, 1-26.

Gilboa, I., L. Samuelson, and D. Schmeidler (2010), "The Dynamics of Induction in a Unified Model", mimeo.

Goodman, N. (1955). *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press.

Hume, D. (1748), *Enquiry into the Human Understanding.* Oxford, Clarendon Press.

Kolodner, J. (1992), "An introduction to case-based reasoning", *Artificial Intelligence Review,* **6**(1), 3-34.

Parzen, E. (1962), "On the Estimation of a Probability Density Function and the Mode", *Annals of Mathematical Statistics*, **33**: 1065-1076.

Riesbeck, C. K. and R. C. Schank (1989), *Inside Case-Based Reasoning.* Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.

Rissland E. L. and Skalak D. B. (1989), Combining case-based and rule-based reasoning: A heuristic approach. Proceedings IJCAI-89, 524-530.

Schank, R. C. (1986), *Explanation Patterns: Understanding Mechanically and Creatively.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Slade, S. (1991), "Case-based reasoning: A research paradigm", *AI Magazine*, 42-55.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis.* London and New York: Chapman and Hall.

Solomonoff, R. (1964), “A Formal Theory of Inductive Inference I, II”, *Information Control*, **7**, 1-22, 224-254.

Wittgenstein, L. (1922), *Tractatus Logico Philosophicus.* London, Routledge and Kegan Paul.