SUBJECTIVITY IN INDUCTIVE INFERENCE[1]

Itzhak Gilboa[2] and Larry Samuelson[3]

December 6, 2010

Abstract

This paper examines circumstances under which subjectivity enhances the effectiveness of inductive reasoning. We consider agents facing a data generating process who are characterized by inference rules that may be purely objective (or data-based) or may incorporate subjective considerations. Agents who invoke no subjective considerations are doomed to ineffective learning. The analysis places no computational or memory limitations on the agents—the role for subjectivity emerges in the presence of unlimited reasoning powers.

[1]We thank Daron Acemoglu, Ken Binmore, Arik Roginsky, the editor and two referees for discussions, comments, and references, and thank the National Science Foundation (SES-0549946 and SES-0850263) for financial support.

[2]HEC, Paris, Tel-Aviv University, and Cowles Foundation, Yale University. tzachigilboa@gmail.com.

[3]Yale University. Larry.Samuelson@yale.edu.

Contents

1 Introduction

1.1 Inductive Inference

Inductive inference is the art of selecting theories based on observations. It is at the heart of scientific and statistical research, as well as much of everyday reasoning. The economist who engages in model selection to explain data, the investor who seeks trends in the behavior of financial markets, and the executive who plans her next marketing campaign all share the same question: Given what I've seen, which rule (or "theory" or "model") should be used to predict future observations?

A first fundamental principle is that one should only consider theories that have *not* been refuted by the data. But how should people choose among the theories that best match the data?

People typically bring subjective criteria to bear in making this choice, tending to select theories that seem *a priori* reasonable, intuitive, simple, elegant, familiar, or that satisfy a variety of other considerations. Why does such subjective reasoning persist? Would it not be better to base model selection on objective criteria alone? Perfectly objective inference is presumably impossible, of course—even the purest of classical statisticians must exercise their judgement when deciding which variables in include in their models—but shouldn't we strive to be as objective as possible?

This paper addresses these questions. We explain how and why subjective criteria are essential to effective reasoning. We conclude that inference cannot effectively be based on likelihood arguments alone—simply observing that one theory fits the data better than another is not sufficient to prefer the former over the latter. Instead, one must also argue that the candidate theory fares well in terms of consistently applied, subjective auxiliary criteria.

1.2 Examples

A pair of examples will set the stage for our argument.

Example 1: Lisa and Sally, adjusting their commuting habits to a new job, check whether the bus runs between their home and the office in consecutive

12-hour periods, observing the pattern

$$0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1,$$

where a 1 denotes that the bus runs. They are now asked their prediction for the seventh and subsequent periods. Lisa sees no objective way for translating these data into a model and hence into a forecast. Anxious to avoid subjective considerations, she randomly chooses one of the many models consistent with the data in each of the subsequent periods, leading to a random choice of 0 or 1. Sally similarly sees no objective way of settling on a model, but nonetheless believes the obvious pattern consistent with these data is "$f(n) = 1$ iff n is even." Who is more likely to be disappointed the next evening, when Sally drives to work while Lisa happens to wait at the bus stop? ■

The basic difficulty is that there is an overwhelming multitude of theories consistent with the data. Lisa's unguided choice from this sea of possibilities is essentially an arbitrary choice, ensuring that she learns nothing from her observations. Sally instead brings some subjective considerations to bear in choosing a model and making a prediction. We have, of course, heightened the contrast by choosing a setting in which Sally's belief seems obvious, while not even raising the question of why this belief is "right." It is accordingly important to note that we identify conditions and a class of subjective beliefs with the property that *any* such belief leads to effective learning, no matter what it is. The key is not that the belief be "correct," but that it allow Sally to escape Lisa's foundering.

Example 2: Lloyd and Sam take a test in which they are asked to extend the sequence

$$1 \quad 2 \quad 4 \quad 8 \quad \ldots .$$

Lloyd (like Lisa) has no way of sorting through the various theories consistent with the data. He realizes that the function

$$f(n) = 2^{n-1} \tag{1}$$

fits the data perfectly well. However, he knows that there are other such functions, such as

$$g(n) = -\frac{1}{3}n^4 + \frac{7}{2}n^3 - \frac{73}{6}n^2 + 18n - 8,$$

that also match the first four observations, in this case predicting that the next observation be 7 rather than 16. In the absence of an objective way to choose among f, g, and many other such functions, Lloyd makes an essentially random choice. By contrast, Sam is comfortable in bringing auxiliary subjective considerations to bear, and finds f more likely to be the correct rule than is g or any other function. Thus, employing some subjective criteria such as simplicity or elegance, Sam predicts that the next observation will be 16. ■

Once again, the set of functions that might have generated the data, and the set of resulting predictions for the next value, is virtually limitless, leading to a prediction that is essentially random, unless some subjective considerations are brought into play. As before, we have designed the example so that Sam's response seems obvious. However, the important point will again turn out to be not that Sam happens to be right (or at least we suspect he is), but rather that he consistently applies *some* appropriately-structured systematic way of sorting theories consistent with the data.

It would be easy to extend the list of such examples, making a compelling case for the pervasiveness of subjective reasoning.

1.3 Results

Our model of inductive inference leads to three conclusions:

- There is no reason to view subjective aspects of inductive inference as shortcomings that push us away from a goal of being as objective as possible. Instead, effective induction requires subjectivity. Inductive inference based on objective criteria alone is bound to fail, while incorporating subjective criteria alongside objective ones can lead to successful learning. Indeed, effective learning requires a willingness to sacrifice goodness-of-fit in return for enhanced subjective appeal.

- Not all subjective criteria are created equal—a subjective criterion will necessarily be effective if and only if it does not treat theories too asymmetrically. Within this class, however, the content of the subjective criterion is much less important. A wide variety of criteria can lead to effective learning.

- Induction will be effective if goodness-of-fit and subjective considerations are balanced so as to produce some stability in the theories used to predict future observations. The history of one's reasoning thus provides a guide as to how one should juggle contending criteria in future reasoning.

We begin in Section 3 with a simple deterministic model that conveys the basic point. Supplementing (objective) likelihood considerations with the consistent application of a subjective ranking of theories dominates relying on objective criteria along. This result rests on a simple enumeration argument: the subjective reasoner will eliminate incorrect theories until she gets to the correct one, thereafter predicting the environment with precision. To be effective, the subjective order must treat theories "not too asymmetrically" in the sense that it allows such an enumeration. In contrast, an agent who relentlessly chases goodness of fit may well never settle on the correct theory, being ultimately doomed to predict no better than chance.

Section 4 extends the results to more realistic settings in which the world about which the agent reasons is random rather than deterministic. Our result that the agent cannot simply rely on goodness-of-fit comparisons is strengthened in this environment. It is an optimal strategy for the agent to regularly reject theories that provide *superior* fits in favor of less successful but subjectively more appealing ones, for much the same reasons that statisticians prefer simpler models and scientists prefer more parsimonious theories in order to avoid the dangers of overfitting their data. To ensure this subjective strategy is successful, however, it must be coupled with a preference for stability. The agent will thus embrace a theory promising enhanced explanatory power only if it is sufficiently subjectively appealing *and* has provided sufficiently good fit for sufficiently long time.

Section 5 discusses extensions and qualifications of our analysis.

This paper complements a vast body of literature in statistics and machine learning that deals with statistical learning.[1] In contrast to this literature, we are interested in optimal learning without assuming that there is an underlying probability law from which the learner can sample in an independent-and-identically-distributed manner. Instead, our main concern

[1]For example,, the Vapnik-Chervonenkis [16, 17] theory, recently applied to decision theory by Al-Najjar [2], deals with the rate at which one can simultaneously learn the probabilities of multiple events.

is the learning of a pattern that has been selected once and for all at the beginning of time. For example, while statistical learning might be concerned with the prediction of the weather on a given day, assuming that it follows an i.i.d. distribution, our main concern would be in determining whether global warming is underway.[2] We are thus interested in a learning problem that is non-trivial even for deterministic processes.

2 The Model

2.1 Overview

We consider a repeated prediction problem. In each period $0, 1, \ldots$, an agent is called upon to predict an observation from the set $\{0, 1\}$. The agent receives a stage payoff of 1 for a correct prediction and 0 for an incorrect one. The agent's objective is to maximize the long-run average of her expected stage payoffs.[3]

The observations are produced by a data generating process, which is simply a function from histories into current observations. The agent's prediction problem would be trivial if she knew the data generating process. We assume she does not, giving rise to a decision problem under uncertainty, where the states of the world correspond to the possible data generating processes.

The agent makes her prediction with the help of theories. A theory, like the data generating process, is a function from all conceivable histories to predictions. The agent has a history-dependent preference relation over theories. In each period, she uses this preference relation to select a theory, which we interpret as her preferred explanation for the history she has observed, which she then uses to make her prediction. This paper studies the basic problem facing the agent, which is to choose the preference relation

[2]There are many economic problems that are appropriately modeled as classical statistical learning, while many others involve too few repetitions to make the i.i.d. sampling a reasonable assumption. For example, one might reasonably describe candidates for admission to a graduate school as a long sequence of i.i.d. (conditional on observable characteristics) repetitions. By contrast, when deciding whether to get married, one has a limited database about oneself and one's prospective spouse. Similarly, predicting whether a particular customer will make a purchase falls under the classical theory of statistical learning, but predicting a stock market crash or a war likely does not.

[3]The assumption that agents are infinitely patient is relaxed in Section 5.3.

over theories that will induce her to make payoff-maximizing predictions.

2.2 Formalities

Observations. At the beginning of each period $n \in \{0, 1, \ldots\}$, the agent observes a profile of variables $x_n = (x_n^1, ..., x_n^m) \in \{0,1\}^m \equiv X$. The agent then predicts the value of another variable, $y_n \in \{0,1\}$, to be revealed at the end of period n. We fix a sequence $\{x_n\}_{n\geq 0}$ and conduct the discussion relative to this sequence, without specifying the process that generated it.[4] Indeed, one could simplify the notation by eliminating the x_n from the model altogether, but we find them helpful for interpretations.

A history of length $n \geq 0$ is a sequence $h_n = ((x_0, y_0), ..., (x_{n-1}, y_{n-1}), x_n)$. The set of all histories of length n is denoted by $H_n = (X \times \{0,1\})^n \times X$. The set of all histories is $H = \cup_{n\geq 0} H_n$, with h denoting an element of H.

The Data Generating Process. A *data generating process* is a function $d : H \to [0,1]$, with $d(h_n)$ being the probability that $y_n = 1$ given history h_n. We let D be the set of possible data generating processes, and hence $D \subset [0,1]^H$. We will often be interested in problems in which the set of possible data generating processes is a strict subset of $[0,1]^H$. For example, we will initially consider the set $\{d \in [0,1]^H |\ d(h) \in \{0,1\}\ \forall h \in H\} \equiv D_0$ of all deterministic data generating processes.

In the course of our discussion, we will consider several possibilities for the set D. It is useful for future reference to collect the notation for these various sets in Figure 1.

Predictions. The agent uses theories to make her predictions. A theory is a function $t : H \to [0,1]$, and hence is simply a candidate data generating function. Which theory the agent uses depends on the history she has observed. This history may tell her that some theories are obviously inapplicable, while suggesting that others are relatively likely to generate correct predictions.

[4]None of our results depends on the characteristics of this data generating process or on realizations of the data having particular properties. In a more general model, some of these variables might be determined by the agent, who might decide to perform experiments and test various theories. Our focus in this paper is on learning without experimentation.

D	Set of data generating processes $\left(\subset [0,1]^H\right)$.
D_0	Set of deterministic data generating processes (i.e., with outputs $\{0,1\}$).
D_0^T	Set of Turing machines with inputs H and outputs $\{0,1\}$.
D_0^H	Set of Turing machines in D_0^T that halt for all $h \in H$.
D_0^B	Set of Turing machines in D_0^T with bounded halting time.
D_ε	Set of data generating processes with outputs $\{\varepsilon, 1-\varepsilon\}$.

Figure 1: Data Generating Processes. In each case, "Set of Turing machines...." should be read "set of data generating process that can be implemented by a Turing machine...."

The basic characteristic of an agent is a collection of relations $\{\succsim_h \subset D \times D;\ h \in H\} \equiv \succsim$ that captures the link between histories and theories. Having reached period n and observed history h_n, the agent uses $\succsim_{h_n}$ to select a theory t_{h_n} from the set D.[5] The agent then uses the theory t_{h_n} to predict the period-n value y_n given history h_n. If $t_{h_n}(h_n) > 0.5$, the agent predicts $y_n = 1$. She predicts $y_n = 0$ if $t_{h_n}(h_n) < 0.5$, and predicts 0 and 1 with equal probability if $t_{h_n}(h_n) = 0.5$.

We assume that, for every h, $\succsim_h$ is complete and transitive, and that it has maximal elements. We define

$$B_{\succsim_h} = \{t \in D \,|\, t \succsim_h t' \quad \forall t' \in D\}$$

to be the set of "best" theories in the eyes of the agent (characterized by $\succsim_h$) faced with history h.

Our interest thus centers on the relation $\succsim_h$. Which specifications of $\succsim_h$ will allow the agent to earn high payoffs?

Payoffs. Given a history h_n, data generating process d and theory t_{h_n}, the probability the next (period-n) prediction is correct is

$$\pi(d, t_{h_n}, h_n) = d(h_n)t_{h_n}(h_n) + (1 - d(h_n))\,(1 - t_{h_n}(h))\,.$$

[5]Sections 5.1 and 5.2 examine the implications of allowing the agent's set of possible theories to differ from the set D of possible data generating processes. Assuming that the agent chooses theories from the set D gives rise to a relatively favorable environment for prediction, by ruling out cases in which the agent cannot conceive of some possible data generating hypotheses (e.g., chooses from a subset of D), and ruling out cases in which she considers theories that are necessarily incorrect (e.g., chooses from a superset of D).

Intuitively, given a data generating process d and a set of theory-selection relations $\succsim$, we would like to take the long-term payoff to the agent to be the limit of the average expected value of these payoffs, or

$$\lim_{T\to\infty} \mathcal{E}\left\{\frac{1}{T}\sum_{n=0}^{T-1}\pi(d, t_{h_n}, h_n)\right\}, \tag{2}$$

where the expectation $\mathcal{E}$ captures the randomness over the histories generated by d and the randomness in selecting theories under $\succsim$. However, this limit need not exist. Let $\Lambda : [0,1]^\infty \to [0,1]$ be a Banach limit defined on the set of infinite sequence of numbers in $[0,1]$. Then we let the agent's payoff $\Pi(d, \succsim)$, when facing data generating process d and using relation $\succsim = \{\succsim_h\}_{h\in H}$ to choose theories, be given by the Banach limit of the resulting sequence of expected values from (2). The key property of Banach limits we need is that

$$\liminf_{T\to\infty} \mathcal{E}\left\{\frac{1}{T}\sum_{n=0}^{T-1}\pi(d, t_{h_n}, h_n)\right\} \leq \Pi(d, \succsim) \leq \limsup_{T\to\infty} \mathcal{E}\left\{\frac{1}{T}\sum_{n=0}^{T-1}\pi(d, t_{h_n}, h_n)\right\},$$

and any payoff criterion with this property would suffice for our results.

2.3 The Likelihood Relation

We described Lisa and Lloyd in our introductory examples as being anxious to use only objective information. This objective information is captured by the *likelihood relation*. The likelihood relation chooses theories that fit the data best. Formally, define the likelihood of theory t given history h_n,

$$L(t, h_n) = \Pi_{j=0}^{n-1}[t(h_j)y_j + (1 - t(h_j))(1 - y_j)].$$

Then the likelihood relation $\succsim^L$ ranks theories after any history h by their likelihood:

$$\forall h \in H, \quad t \succsim_h^L t' \quad \iff \quad L(t, h) \geq L(t', h).$$

The likelihood relation thus calls for agents to base their inferences on their data, and on no other criterion.

In the simplest case, when only deterministic theories are considered, $\succsim_h^L$ boils down to two equivalence classes. All theories that perfectly fit the data are equivalent, having $L(t, h) = 1$, and they are all preferred to all theories that have been refuted by the data, where the latter are also equivalent to each other and satisfy $L(t, h) = 0$.

3 Deterministic Data Processes: Subjectivity in Inductive Inference

This section uses an elementary deterministic model to show how subjective criteria can be useful in inductive inference. The key restriction in the model is contained in the following assumption, which puts some structure on the data generating processes. Its first two parts require the set of data generating processes be simple enough to be learned, and the final part requires it be rich enough to describe any possible finite sequence of observations. The latter is intended to rule out trivial cases in which a finite set of observations suffices to single out a unique theory, i.e., cases where the problem of induction does not arise.

Assumption 1

[1.1] $D \subset D_0$*, the set of deterministic data generating processes.*

[1.2] D *is countable.*

[1.3] *For every history* $h \in H$ *there exists* $d \in D$ *such that* $L(d, h) = 1$*.*

Given Assumption 1.1, the remaining requirements are satisfied if D is the set D_0^H of all Turing machines generating functions $d \in D_0$ (i.e., D_0^H is the set of Turing machines that accept elements of the set H as inputs, halt, and produce outputs from the set $\{0, 1\}$). The countability restriction will be discussed and relaxed in Section 5.1 below.

3.1 Applying the Likelihood Relation

We now consider the performance of an agent who consistently applies the likelihood relation $\succsim_h^L$ as a guide to making predictions, but applies no other criteria.

The agent's choice of theory when using a relation $\succsim_h$ and following history h is unambiguous if $B_{\succsim_h}$ is a singleton, but this may often fail to be the case. What does the agent do if there are a number of theories in $B_{\succsim_h}$? We assume that the agent treats the various best theories symmetrically, in the sense that she makes a choice that wherever possible exhibits no bias for theories that predict 0 versus theories that predict 1 in the next observation.

To make this assumption precise, notice that the set D_0^H of all Turing machines generating functions $d \in D_0$ has the property that for any history of observations h, and for every data generating process d in D_0^H consistent

with h, there is another data generating process in D_0^H that is also consistent with h but whose subsequent datum will be precisely the opposite of d, generating a 0 whenever d produces a 1 and vice versa. As a result, the sets $\left\{t \in B_{\succsim_h^L} \,|\, t(h) = 0\right\}$ and $\left\{t \in B_{\succsim_h^L} \,|\, t(h) = 1\right\}$ will not only be non-empty but will be symmetric in their treatment of the next observation. The obvious implementation of our unbiased-choice provision is then:

Assumption 2 *The agent chooses a theory from $B_{\succsim_h}$ according to a measure $\mu_{B_{\succsim_h}}$ on $B_{\succsim_h}$ satisfying*

$$\mu_{B_{\succsim_h}}\left(\left\{t \in B_{\succsim_h} \,|\, t(h) < 0.5\right\}\right) = \mu_{B_{\succsim_h}}\left(\left\{t \in B_{\succsim_h} \,|\, t(h) > 0.5\right\}\right)$$

whenever

$$\left\{t \in B_{\succsim_h} \,|\, t(h) < 0.5\right\}, \left\{t \in B_{\succsim_h} \,|\, t(h) > 0.5\right\} \neq \varnothing.$$

We then have:

Proposition 1 *Let Assumptions 1 and 2 hold. Then $\Pi(d, \succsim^L) = \frac{1}{2}$.*

The proof, contained in Section 6, is built on the following observations. There are always many theories consistent with whatever data the agent has observed. In particular, after every history, the set of unfalsified theories available to the agent contains theories that predict a 0 as well as theories that predict a 1, and leave the agent with no means of choosing between the two sets of theories. The agent's choice is thus random, ensuring a long-run payoff of 1/2. Without some means of eliminating theories, the agent can thus never predict better than chance. Unfortunately, the data alone provide no possibilities for such elimination.

3.2 The Subjective Relation

We now consider an agent who, like Sally and Sam in our examples, brings subjective criteria to bear in choosing between theories. To define such a theory-selection procedure, we begin with an order $\succ^S \subset D \times D$ that is defined *a priori*, independently of history (and hence is "subjective," in contrast to

$\succsim_h$ which is a function of $h \in H$).[6] We require $\succcurlyeq^S$ to be complete and transitive. In addition, we say that $\succcurlyeq^S$ is a *discriminating* subjective order if

$$\#\left\{t' \in D \,|\, t' \succcurlyeq^S t\right\} < \infty \qquad \forall t \in D. \tag{3}$$

Condition (3) has two important implications. Most obviously, it ensures that the subjective order's indifference classes are not too large. The problem with the likelihood relation is that it leaves the agent with too many indifferences, in the sense that the agent will be stuck choosing among too many theories that fit the data. The subjective order will help select between such indifferent theories, but will be effective only if it does a good enough job of breaking indifferences. In the absence of condition (3), for example, the definition of a subjective order would be consistent with the trivial order $\succcurlyeq^S = D \times D$, according to which no theory is ranked ahead of another, giving the agent absolutely no help in choosing between theories. More generally, (3) rules out cases in which the subjective order is permissive enough to allow for infinitely many strategies to be grouped in a single indifference class. However, (3) does much more than simply limit indifference classes, as we make clear in Section 3.3.

One natural way to ensure that (3) holds is to enumerate D and set $t_i \succ^S t_{i+1}$ for every $i \geq 1$. Our condition is less demanding, and allows for non-singleton equivalence classes of the order $\sim^S$, but not for infinite ones. Nonetheless, under the assumption that D is countable, discriminating subjective orders are closely related to enumerations of D. Specifically, for every discriminating order $\succcurlyeq^S$ there exists an enumeration $D = \{t_1, t_2, ...\}$ such that $t_i \succcurlyeq^S t_{i+1}$, with strict preference $\succ^S$ occurring for infinitely many i's. Alternatively, $\succcurlyeq^S$ is a discriminating subjective order if and only if it can be represented by a function $C : D \to \mathbb{N}$ such that[7]

$$t \succcurlyeq^S t' \quad \iff \quad C(t) \leq C(t') \tag{4}$$

[6] In an effort to keep things straight, we use $\succsim$ to denote a relation by which the agent chooses theories, and $\succcurlyeq$ to denote a subjective order over theories. We similarly associate the label "relation" with the former and "order" with the latter (though they have the same properties, i.e., each is complete and transitive).

[7] While there are obviously many different enumerations of D, and hence many functions C with their induced orders $\succcurlyeq^S$, they cannot be too different in the following sense. Let C_1 and C_2 be two such functions. Then, for every k there exists $l = l(k)$ such that $C_1(t) > l$ implies $C_2(t) > k$. That is, a theory that has a sufficiently high index according to C_1 will also have a high index according to C_2.

and

$$|C^{-1}(k)| < \infty \qquad \forall k \in \mathbb{N}.$$

Given a subjective order $\succcurlyeq^S$, we define the associated subjective relation $\succsim^{LS}$ for choosing theories as follows:

$$\forall h \in H, \quad t \succsim_h^{LS} t' \iff \begin{cases} \{t \succ_h^L t'\} \\ or \quad \{t \sim_h^L t' \ \ and \ \ t \succcurlyeq^S t'\} \end{cases}.$$

The relation $\succsim^{LS}$ thus uses the subjective order $\succcurlyeq^S$ to choose among those theories with the highest likelihood. The likelihood and subjective relations $\succsim^L$ and $\succsim^{LS}$ agree in that they only choose theories with maximal likelihoods, with the likelihood relation being indifferent over such theories and the subjective relation providing the criterion for making this choice.

A discriminating subjective order may still frequently render the agent indifferent over many theories. The following result does not depend upon how these indifferences are broken, and hence requires no counterpart of Assumption 2.

Proposition 2 *Let Assumption 1 hold. For every discriminating subjective order $\succcurlyeq^S$ and every $d \in D$, $\Pi(d, \succsim^{LS}) = 1$. Hence, for a discriminating subjective order $\succcurlyeq^S$, the induced subjective relation $\succsim^{LS}$ strictly dominates the likelihood relation $\succsim^L$.*

The agent begins the prediction process with no data, and accordingly chooses from the first indifference class in her subjective order. This indifference class may not be a singleton, and she may shift among the theories in the class as data accumulate, even if none are falsified. However, the only event that will push her outside this indifference class is for each of its elements to be falsified by the data. Moreover, this first indifference class is finite. If it contains the actual data generating process, the agent will never be pushed out of this class, and will eventually be limited to a collection of theories that are observationally equivalent to the data generating process, ensuring correct predictions. Alternatively, if this indifference class contains neither the actual data generating process nor any observationally equivalent process, the agent will eventually be pushed into her second indifference class. Here, we can repeat the same reasoning, continuing until the agent settles on a theory that makes correct predictions.

3.3 Which Subjective Relations Work?

3.3.1 Exploitation and Exploration

What makes the subjective relation work, and what stymies the likelihood relation? The subjective relation allows effective prediction because it embodies the principles of "exploitation and exploration." The agent exploits theories that have worked by sticking with them, while effectively exploring new theories when necessary. The persistent appeal to the agents' subjective order, whatever the order might be, ensures that a theory that fits the data is not abandoned, while the enumeration provided by the order ensures that the agent will "try out" all theories (as long as a perfect fit has not been found). The likelihood relation's lack of the first characteristic dooms its adherents to randomness.

3.3.2 Two Examples

The assumption that the subjective order is discriminating plays a role in ensuring both exploration and exploitation. We illustrate with two examples.

Example 3: Consider a subjective order that ranks any theory predicting an initial 0 in a single indifference class that comes ahead of all others, and then enumerates the remaining theories and (strictly) ranks them accordingly. Suppose the actual data generating process produces a 0 in the first period. Then the agent will never be pushed beyond her first indifference class. In addition, the subjective relation provides no guidance as to how the agent should choose from this topmost indifference class, leaving the agent in the same random-choice predicament as does the likelihood relation. In this case, the subjective order does not ensure adequate exploitation. The most obvious purpose of (3), noted just after its introduction, is to preclude such cases.

Example 4: Suppose the subjective order is based on a lexicographic accounting of 1s. Theory t is ranked ahead of t' if t predicts a 1 in the first period and t' a 0. If they make the same first-period prediction, then t is ranked ahead of t' if t predicts a 1 in the second period and t' a 0. If they make the same predictions in the first two periods, then t is ranked ahead of t' if t predicts a 1 in the third period and t' a 0, and so on. No two theories are indifferent under this order, so that exploiting a theory corresponding

to the actual data generating process, once one has reached it, is assured. Suppose, however, the data generating process produces a perpetual string of 0s. The theory corresponding to this outcome ranks below every other possible theory. The agent will never reach this theory, and indeed will predict a 1 in every period, earning a payoff of 0 that makes random choice look inspired. In this case, it is exploration that is lacking. The second key aspect of (3) is to preclude such possibilities. For every possible theory, there are only finitely many preferred theories under $\succ^S$, ensuring that exploration guided by $\succ^S$ will eventually hit upon the data generating process (or something observationally equivalent), at which point this theory will be effectively exploited.

3.3.3 An Inertial Likelihood Relation

Suppose we build more effective exploitation into the likelihood relation by assuming that agents do not abandon a theory until receiving evidence of its falsity. In particular, the proof of Proposition 1 shows that an agent guided by the likelihood relation falters because every period there is a multitude of theories with perfect likelihood scores, including the truth and a host of imposters. The agent's arbitrary choice from this set implies that even if she hits upon the truth, she soon abandons it in favor of another seemingly equivalent theory. Will it not suffice to assume that the agent sticks to something that has worked in the past?

The phenomenon of inertia, or a preference for a status quo, is familiar from casual observations as well as from psychological studies. Kuhn [9] argued that scientists tend to cling to old theories rather than adopt those theories that fit the data best. More recently, we see traces of inertia in the status-quo preference used in behavioral economics. Indeed, we might think of inertia as another subjective consideration used to supplement the likelihood relation.

To see if inertia suffices for effective learning, we define the *inertial* relation as that selecting the theory chosen in the previous period if the latter maximizes the likelihood function, and otherwise choosing as does the likelihood relation. Formally, define $\succsim^{LI}$ as follows for all $n > 1$,

$$\forall h \in H, \qquad t \succsim_h^{LI} t' \iff \begin{cases} & \{L(t,h) > L(t',h)\} \\ \textit{or} & \{L(t,h) = L(t',h) \ \textit{ and } \ t = t_{n-1}\} \\ \textit{or} & \{L(t,h) = L(t',h) \ \textit{ and } \ t, t' \neq t_{n-1}\} \end{cases},$$

with $t \sim_{h_0}^{LI} t'$ for all t, t', so that in the absence of any evidence, all theories are equally likely.

The following example shows that inertia alone does not suffice to ensure effective learning.

Example 5: Let (for this example only) D consist of the following set of deterministic theories:

$$\left\{ y \in \{0,1\}^{\mathbb{N}} \mid \exists n \geq 0, \quad y(k) = 0 \ \forall k \geq n \right\}.$$

The possible data generating processes are thus all those that generate only 0 from some point on. For example, the theories may be describing the availability of a random resource, which is known to be depletable, but whose date of ultimate exhaustion is uncertain.

For every h_n, let the selection rule over the infinite set $B_{\succsim_{h_n}^{LI}}$ be given by

$$\mu_{B_{\succsim_{h_n}^{LI}}}\left(t^{n+k}\right) = \frac{1}{2^{k+1}} \qquad k = 0, \ldots, \tag{5}$$

where, for all histories and all k,

$$t^{n+k}(h_{n+k}) = 1; \quad t^{n+k}(h_l) = 0 \quad \forall l \neq n+k. \tag{6}$$

Hence, given any history h_n, the agent attaches positive probability only to continuations that feature a single observation of 1 (and otherwise all 0's), with probability $1/2^{k+1}$ attached to the theory that generates its observation of a 1 in precisely k periods.

Under this selection rule, theories predicting a 0 on the next step are equally likely as theories predicting a 1, in accordance with (2). Suppose the data generating process is such that $y_n = 0$ for all n. Consider $\succsim_{h_n}^{LI}$ for a history h_n consisting of n 0s. Then given (5)–(6), $\succsim^{LI}$ will choose a theory whose first 1 appears according to a geometric distribution with parameter 0.5. The expected number of periods for which this theory will match the observations is

$$\sum_{i=0}^{\infty} \frac{i}{2^{i+1}} = 1.$$

It is then a straightforward calculation that $\Pi(y_0, \succsim^{LI}) = 1/2$.[8] ■

The difficulty in this example is that the selection rule over the various sets $B_{\succsim^{LI}_{h_n}}$ routinely ignores the correct theory. Exploitation is assured, but exploration is not. Proposition (3) shows that inertia can be valuable, in effect serving as a safeguard against the excessive fickleness of random choice, if we also take steps to ensure effective exploration.

Assumption 3 *There exists a strictly positive measure λ on the countable set D such that for any $h \in H$, $\mu_{B_{\succsim_h}}$ equals λ conditioned on $B_{\succsim_h}$.*

Proposition 3 *Under Assumptions 1 and 3, for all $d \in D$, $\Pi(d, \succsim^{LI}_h) = 1$.*

Behind this result lies the observation that, under Assumption 3, the theory selection process is guaranteed to select the correct theory, d, at least once. Once d has been chosen, inertia ensures that it will not be abandoned, and hence the optimal payoff is obtained.

3.3.4 Bayesian

If the set of conceivable data generating processes is countable (cf. Assumption 1) and the agent has a Bayesian prior over this set, then the relation "has at least as high a prior as" can be viewed as a subjective order—it is a weak order that is monotonically decreasing along an enumeration of the theories, with finite equivalence classes. In other words, a Bayesian prior defines a subjective order. Conversely, one may use a subjective relation to define a Bayesian prior: theories ranked higher under the subjective order are considered more likely.

There are a continuum of priors that are consistent with a given subjective order. These priors are all equivalent in our model, because we suggest that the agent choose a most-likely theory to generate the next prediction. By contrast, the Bayesian approach constructs an expected prediction, using all possible theories.

[8] We can write the agent's payoff, at either the beginning of the learning process or upon having had a theory falsified and choosing a new one, as $V = \lim_{T\to\infty} \sum_{i=1}^{T-1} (p_i(i-1) + V)$, where p_i is the probability that the first 1 under the newly chosen theory occurs in the ith period. This gives $V = \lim_{T\to\infty} \sum_{i=0}^{T-1} \left(\frac{1}{2^i}(i-1) + V\right)$ and hence $V = \frac{1}{2}$.

Observe that the Bayesian approach is cognitively rather demanding, because it requires a quantification of the relative likelihood of all theories, and therefore also a priori awareness of all theories, whereas the selection of a single theory for prediction does not require the formulation of alternative theories that will be used once that theory fails. However, in terms of limit of the payoff, the selection of a single theory suffices for effective learning.[9]

4 Random Data Generating Processes: Likelihood Tradeoffs

The assumption that the data generating process is deterministic (i.e., that $d(h) \in \{0,1\}$ for all h) is unrealistic. Worse still, it beclouds the interesting trade-off between likelihood and subjective considerations in the choice of theories. So far, the choice of theories was made among the theories that fit the data perfectly, and thus subjective idiosyncracies involved no cost. But when random data generating processes are introduced, subjective considerations are no longer a free good, but impose a price in terms of likelihood. Should the agent be willing to give up a better fit for a subjectively more appealing theory, and if so, to what extent?

4.1 Uniform Errors

To get some insight into this problem, we begin with a minimal modification of our benchmark model. Define, for $\varepsilon \in (0, 1/2)$,

$$D_\varepsilon = \{d \in [0,1]^H \,|\, d(h) \in \{\varepsilon, 1-\varepsilon\} \ \forall h \in H\}.$$

Thus, D_ε can be thought of as the deterministic data generating processes, D_0, with an error probability of ε added to the output.

The likelihood function, for a theory $t \in D_\varepsilon$ and a history $h \in H_n$, is

$$L(t, h_n) = \Pi_{j=0}^{n-1}(t(h_j)y_j + (1 - t(h_j))(1 - y_j)).$$

In the presence of randomness, the likelihood function will inevitably converge to zero for any theory: its largest possible value in period n is $(1-\varepsilon)^n$,

[9] If we are interested in discounted payoffs, then it is relevant to note that the Bayesian approach will never obtain perfect prediction, because it will always entertain beliefs that are wrong.

$$\begin{array}{rcrl} \log(1-\varepsilon) & = & \theta(1) & \text{Maximum possible limiting value.} \\ (1-\varepsilon)\log(1-\varepsilon)+\varepsilon\log\varepsilon & = & \theta(1-\varepsilon) & \text{Value achieved by the data generating process.} \\ \frac{1}{2}\log(1-\varepsilon)+\frac{1}{2}\log\varepsilon & = & \theta\left(\frac{1}{2}\right) & \text{Value achieved by random choice.} \end{array}$$

Figure 2: Key values of the limiting average-log-likelihood function (7).

since the best any theory can do is attach probability $1-\varepsilon$ in each period to the outcome that happened to be realized in that period. This convergence makes the likelihood an awkward standard for comparing theories. It is more convenient to consider the average of the logarithm of the likelihood function,

$$\begin{aligned} l\left(t,h_n\right) &= \frac{1}{n}\log(L(t,h_n)) \\ &= \frac{1}{n}\sum_{j=0}^{n-1}\log\left[t(h_j)y_j+(1-t(h_j))(1-y_j)\right], \end{aligned} \tag{7}$$

which does not converge to zero. We hereafter use "likelihood" to denote the average log likelihood, given by (7).

Let us say that a theory is "correct" in period t if it predicts a 1 with probability $1-\varepsilon$ and a 1 occurs, or if it predicts a 0 with probability $1-\varepsilon$ and a 0 occurs. It is helpful to define the function

$$\theta(p)=p\log(1-\varepsilon)+(1-p)\log\varepsilon.$$

Then $\theta(p)$ is the (average log) likelihood of a theory that has been correct proportion p of the time.

A theory that is correct in every period would give likelihood $\theta(1)$. This is the highest possible likelihood. The theory that corresponds to the data generating process gives a limiting likelihood of $\theta(1-\varepsilon)$, and an agent who always uses the data generating process to predict would achieve payoff $1-\varepsilon$.[10] Predicting randomly would give likelihood $\theta\left(\frac{1}{2}\right)$ and payoff $\frac{1}{2}$. Figure 2 summarizes these observations.

The counterpart of Assumption 1 is now:

Assumption 4

[4.1] $D\subset D_\varepsilon$.

[4.2] *D is countable.*

[4.3] *For every history $h\in H$ there exists $d\in D$ such that $l(d,h)=\theta(1)$.*

[10]For large n, the likelihood will be approximately$(1-\varepsilon)^{(1-\varepsilon)n}\varepsilon^{\varepsilon n}$ and the average log likelihood $l(d,h)$ will converge to $\theta(1-\varepsilon)$.

Assumption 4.3 indicates that for any finite stream of data, there is a theory that would have been correct in every period. Ex post, one can rationalize anything.

4.2 Tolerance in Learning

The agent could once again adopt a relation over theories that first restricts attention to likelihood-maximizing theories, such as the likelihood relation $\succsim^L$ of Section 2.3 or the subjective relation $\succsim^{LS}$ of Section 3.2. In the random environment, this ensures that the agent will eventually *exclude* the data generating process as a possible theory. In each period, the realization may differ from the true theory's prediction with probability ε. Hence, the true theory will eventually almost surely have a likelihood value lower than $\theta(1)$, whereas there will always be other theories with a likelihood value of $\theta(1)$. That is, insisting on maximum-likelihood theories will lead to constant theory hopping.

This suggests that the agent's learning might be more effective if it incorporates some tolerance for inaccuracy. For any $\gamma \in [0,1]$, we say that a theory t is a "γ-best fit" to the data after history h if

$$l\,(t,h) \geq \theta(\gamma).$$

The counterpart of the likelihood relation is then

$$\forall h \in H, \quad t \succsim_h^{L,\gamma} t' \quad \Longleftrightarrow \quad L^\gamma(t,h) \geq L^\gamma(t',h)$$

where

$$L^\gamma(t,h) = \min\{L(t,h), \theta(\gamma)\}.$$

When working with D_0, the likelihood relation $\succsim^L$ separated theories into two classes, those that predicted perfectly and those that did not. The key characteristic of the relation $\succsim_h^{L,\gamma}$ is that it allows us to group the theories achieving a likelihood of at least $\theta(\gamma)$ into a single equivalence class.

What would be a good value of γ? One suspects that we should set $\gamma < 1-\varepsilon$, since any value $\gamma > 1-\varepsilon$ will eventually surely exclude the true data generating process. However, simply relaxing the likelihood threshold to $\gamma < 1-\varepsilon$ does not suffice if one insists on using the likelihood criterion alone to choose theories. The true theory (if such exists) will not be ruled out, but there is no guarantee that it be selected. An argument analogous to that establishing Proposition 1 immediately provides the (omitted) proof of:

Proposition 4 *Let Assumptions 2 and 4 hold. Then* $\Pi(d, \succsim^{L,\gamma}) = \frac{1}{2}$.

Intuitively, whatever the value of γ, the agent has a wealth of theories with likelihoods exceeding $\theta(\gamma)$ from which to choose. In the absence of another selection criterion, the agent is doomed to random prediction.

Once the agent is willing to pay the price of less than maximum likelihood, she can afford to use an additional subjective criterion in a meaningful way. Define

$$\forall h \in H, \quad t \succsim_h^{LS,\gamma} t' \iff \begin{cases} & \{t \succ_h^{L,\gamma} t'\} \\ \text{or} & \{t \sim_h^{L,\gamma} t' \ \text{ and } \ t \succcurlyeq^S t'\} \end{cases}.$$

The agent thus chooses the subjective order to choose among the γ-best fits.

Under the subjective relation, setting $\gamma > 1 - \varepsilon$ again implies that the agent will discard the data generating process as a possible theory and subsequently hop between imposters. The implications of this switching between strategies are now not completely obvious. The agent uses here subjective criteria to choose among the γ-best-fit theories. While the correct theory is not among them, it is not clear how well their predictions are correlated with the true data generating process. The following assumption ensures that the top-rated theories in the subjective order are rich enough to contain theories that predict 0 and theories that predict 1.

Assumption 5 *For a subjective order* $\succsim^{LS,\gamma}$ *with* $\gamma > 1 - \varepsilon$ *and sufficiently large* n, $\left\{t \in B_{\succsim_h^{LS,\gamma}} \,|\, t(h) = 1 - \varepsilon\right\}$ *and* $\left\{t \in B_{\succsim_h^{LS,\gamma}} \,|\, t(h_n) = \varepsilon\right\}$ *are nonempty.*

It is not obvious that the subjective relation should have this property. If, for example, we observe the pattern 00000, it may not be that one of the theories ranked highest by the subjective order will predict 1. However, when n is large, the actual data generating process has surely been discarded by the order $\succsim_h^{LS,\gamma}$ and any theory amassing a likelihood above γ is surely a fluke. As a result, it is not clear what *a priori* information, if any, should be brought to bear, in which case Assumption 5 may be reasonable. The (omitted) proof of the following is then immediate:

Proposition 5 *Let Assumptions 2, 4.1–4.2 and 5 hold. Let* $\gamma > 1 - \varepsilon$. *Then* $\Pi(d, \succsim^{LS,\gamma}) = \frac{1}{2}$.

The key point is that setting $\gamma > 1 - \varepsilon$ forces the agent to abandon any theory that sufficiently often predicts as does the true theory, in the process placing constraints on the payoff of which the agent can be assured. Assumption 5 makes these constraints precise, dooming the agent to random choice.

4.3 Stability in Learning

One virtue of a subjective order in a deterministic environment is that it prevents the agent from abandoning perfectly good theories. Setting $\gamma < 1-\varepsilon$ ensures that the data generating process will at least eventually be among the γ-best fits considered by the agent. This alone, however, does not ensure effective learning. Selecting the subjectively best of the γ-best fit leaves open the possibility that the agent may switch back and forth between theories, where, at each period, one of the theories provides a γ-best fit, but fails to predict correctly. This is possible if the subjective order selects theories that tend to be wrong precisely when they are used for prediction, but "catch up" in terms of the likelihood during periods in which they are not used for prediction. To see that this learner's nightmare might come true, consider the following.

Example 6 Fix $\gamma < (1 - \varepsilon)$ and let d be the data generating process. To simplify the presentation, but without losing any generality, assume that d predicts 1 in each period (with probability $1 - \varepsilon$).

We construct k theories, denoted by $t_1, \ldots, t_k$, which will be ranked at the top of the subjective order: $t_1 \succ^S t_2 \succ^S \ldots \succ^S t_k$ and $t_k \succ^S t'$ for all $t' \notin \{t_1, ..., t_k\}$.

For concreteness, we describe the theories by an algorithm. For $n = 0$, $t_i(h_0) = 1$ for all $i \leq k$. For $n > 0$, given history h_n, every t_i $(i \leq k)$ computes the predictions of all t_j $(j \leq k, j = i$ included) for all sub-histories h_m of h_n (for all $m < n$). By induction, this is a computable task. Next, each t_i computes $l(t_j, h_n)$ for all $j \leq k$. If none of them has a likelihood $l(t_j, h_n) \geq \gamma$, t_i predicts 1. Otherwise, t_i finds the best (under $\succ^S$) of the theories in $\{t_1, ..., t_k\}$ with $l(t_j, h_n) \geq \gamma$. If it is itself, it predicts 0; otherwise, it predicts 1.

Observe that each theory $\{t_1, ..., t_k\}$ basically performs the same algorithm, which simulates the calculations of all previous periods, and halts by induction. The difference between the predictions of the different theories in

$\{t_1, ..., t_k\}$ arises only out of the very last step of the algorithm, in case some of them obtain a likelihood value above the threshold.

Observe also that in each period, at least $k-1$ of the theories $(t_1, \ldots, t_k)$ will produce a prediction matching that of d, and—if and only if some reach the appropriate likelihood threshold—one of these theories will dissent. Let ε_n be the proportion of realized 0's up to time n. The collective number of correct predictions among the k theories $(t_1, \ldots, t_k)$ in history h_n will thus be at least

$$[(1-\varepsilon_n)(k-1)]n,$$

where ε_n gets arbitrarily close to ε with arbitrarily large probability as n gets large. Hence, a lower bound on the number of correct predictions, among the k theories $(t_1, \ldots, t_k)$ over periods $0, \ldots, n-1$ is given by

$$[(1-\varepsilon-\delta)(k-1)]n$$

for some $\delta > 0$. We can choose n^* sufficiently large that

$$\delta < \frac{(1-\varepsilon)-\gamma}{2}$$

and then k sufficiently large that, for all $n > n^*$,

$$\left[(1-\varepsilon-\frac{(1-\varepsilon)-\gamma}{2})(k-1)\right]n > k\gamma n, \tag{8}$$

or

$$\frac{k-1}{k}\left(\frac{1-\varepsilon+\gamma}{2}\right) > \gamma.$$

(Since $1-\varepsilon > \gamma$, such a k exists.) From (8), we see that the theories $(t_1, \ldots, t_k)$ must have collectively amassed at least $k\gamma n$ correct predictions for any $n > n^*$, ensuring that at least one of them must have at least γn correct predictions, and hence a likelihood of at least $\theta(\gamma)$. As a result, one of these theories will be used for prediction in every period $n > n^*$, and by definition predicts that outcome which appears with probability ε under the data generating process d. Hence, the agent's payoff converges to ε. ■

It may appear as if the theories $(t_1, \ldots, t_k)$ in Example 4 are hopelessly special, tied closely to the structure of the true data generating process, and hence that the example is simply a curiosity. While we make no claims for

the realism of the example, it is important to note that the subjective order may rank a multitude of collections $(t_1, \ldots, t_k)$ at the top, each allowing an outcome of the type we have just described for a particular specification of the data generating processes. For any fixed data generating process d, the likelihoods of those theories ranked highly by the subjective order that do not correspond to d will fall until they are irrelevant. The calculations of the example will then become relevant. While still delicate, the phenomenon in the example is thus not as special as it may first appear. If we are to achieve a general result, we must have some additional structure.

The key to addressing this difficulty is to rely on theories that have been *consistently* successful at explaining the data, rather than theories that boast a great likelihood only at the present moment. Formally, let there be given $\gamma \leq 1 - \varepsilon$ and $k \geq 1$. For a theory t and history $h \in H_n$, $n \geq k$, define

$$\Gamma_{\gamma,k}(t) = \sum_{j=k}^{n} \delta_j,$$

where

$$\delta_j = \begin{cases} 1 & if \quad l\,(t, h_j) \geq \theta(\gamma) \\ 0 & if \quad l(t, h_j) < \theta(\gamma) \end{cases}$$

(where h_j is the j-th prefix of h). Next, define the relations $\succsim_h^{LS,\gamma k}$ for $h \in H$ as follows:

$$t \succsim_h^{LS,\gamma k} t' \iff \begin{cases} & [\Gamma_{\gamma,k}(t) > \Gamma_{\gamma,k}(t')] \\ or & [\Gamma_{\gamma,k}(t) = \Gamma_{\gamma,k}(t') \quad and \quad t \succcurlyeq^S t']. \end{cases}$$

Thus, a maximizer of $\succsim_h^{LS,\gamma k}$ has to be a theory that has obtained an average log-likelihood of at least $\theta(\gamma)$ as often as possible over the past consecutive $(n - k + 1)$ periods. If there are several theories that obtained this likelihood threshold for the entire period, the maximizer has to be one that is ranked topmost by the subjective order. If no theory has done as well as $\theta(\gamma)$ for $(n - k + 1)$ periods (perhaps because $k > n$)), $\succsim_h^{LS,\gamma k}$ selects the subjectively-best-ranked among those that have achieved at least $\theta(\gamma)$ for at least $(n - k)$ periods out of the past $(n - k + 1)$ periods, and so forth.

Clearly, the choice of the parameters γ and k allows a wide range of relations $\left(\succsim_h^{LS,\gamma k}\right)$. What should be the values of γ and k and how are they

determined?[11] In particular, different values of ε will call for different values of γ and k.

4.4 Optimal Tolerance

Suppose first that γ and k are selected *a priori*, either as a deliberate choice on the part of the agent or as the result of an evolutionary process that favors effective values of the tolerance for accuracy γ and the taste for stability k, at the expense of ineffective values. How much inaccuracy should the reasoner be willing to tolerate? The critical value $1-\varepsilon$ builds sufficient tolerance for inaccuracy into the agent's choices as to ensure effective learning:

Proposition 6 *Under Assumption 4, for every discriminating subjective order* $\succcurlyeq^S$ *and for every* $d \in D$, $\Pi(d, \succsim^{LS,\gamma^k}) \to (1-\varepsilon)$ *as* $\gamma \nearrow 1-\varepsilon$ *and* $k \to \infty$.

We thus find that, in the presence of randomness, augmenting the subjective order with a preference for stability again enhances the agent's payoff. The argument is quite similar to that of Proposition 2. There are only finitely many theories ranked ahead of the true data generating process d under the discriminating subjective relation. Setting $\gamma < 1-\varepsilon$ ensures that likelihood considerations do not exclude d, while pushing γ quite close to $1-\varepsilon$ ensures that theories whose predictions are quite close to those of d but nonetheless different are eventually excluded. Finally, as k becomes large, the chances that one of the theories ranked ahead of d by the subjective order can have predicted as well as d over any string of k periods becomes negligible. This ensures that the agent will eventually use a theory other than d to make predictions only in extraordinarily rare circumstances, giving the result.

We view this result as intuitive. We tend to trust experts who have *always* provided good explanations more than experts who have *sometimes* provided good explanations. Even if two experts, or theories, reach the same level of goodness of fit at present, a better history may well be a reason to prefer one over the other.

Observe that one cannot do away with the subjective order and rely on stability alone. In the absence of the subjective order, for every history h_n there exists a theory t_n such that $l(t_n, h_j) = \theta(1)$ for *every* $j \leq n$. Such a theory would maximize the likelihood function for each prefix of the history

[11] Notice that it makes no sense to insist on stability if one sets $\gamma > 1-\varepsilon$, since we know that no theory can long sustain a likelihood above $1-\varepsilon$.

h_n, and would therefore be chosen for prediction. Thus the preference for stability alone does not provide a safeguard against overfitting the data by choosing a theory post-hoc.

4.5 Endogenously Determined Tolerance

Proposition 6 suggests that for effective decision making, the optimal tolerance level γ must be large, but not too large. Can we expect the agent to hit upon the optimal level a priori? Instead, the agent need not do so. The agent's decisions provide the information required to ascertain an effective value of the tolerance level γ.

Proposition 7 *Let Assumption 4 hold. For every discriminating subjective order $\succ^S$ there exists a relation $\succsim^{S*}$, independent of ε, such that*

(i) for every $d \in D$, we have $\Pi(d, \succsim^{S*}) = 1 - \varepsilon$

and

(ii) for every $t, t' \in D$, for large enough n, if $\Gamma_{\gamma,k}(t) = \Gamma_{\gamma,k}(t')$, then

$$t \succsim^{S*} t' \qquad \Longleftrightarrow \qquad t \succ^S t'.$$

The idea behind this result is that an agent who sets a level of γ too high will soon find herself switching frequently between theories. This switching can serve as a signal to the agent that she needs to reduce γ. The relation $\succsim^{S*}$ essentially adjusts γ in response to such signals until finding the boundary at which higher values of γ lead to volatile theory choices. The relation then exploits this boundary level of γ much as would an agent who sets the boundary value a priori and implements $\succsim^{LS,\gamma k}$.

The agent who implements $\succsim^{S*}$ engages not only in learning but also in meta-learning. This agent selects theories that provide a γ-best fit and that fare well under the subjective order, but at the same time, she observes her own learning process and learns from this process itself. Specifically, the agent looks at the choices she would have made for various levels of γ and asks, "What can I learn from the fact that for some levels of γ my learning process would have continued indefinitely, whereas for others I would have settled on a specific theory?" The fact that a certain level of γ does not let the agent converge on a given theory is taken to be an indication that this level is too high.

The parameter γ may be viewed as the agent's aspiration level for the degree of accuracy of the theory (in the sense of Simon [13]). We can imagine the agent setting a large value of γ in the hope of finding a theory that is quite close to the maximal likelihood one. However, if she finds that the search for such a theory does not result in a stable choice, and that she keeps bouncing around among theories no matter how large n is, then the agent may reduce her aspiration level γ. When γ is low enough, the agent will find a theory that has a higher degree of inaccuracy, but that can be chosen over and over again. This search for the optimal γ can be viewed as the search for the optimal aspiration level.

4.6 More General Error Specifications

The arguments behind Propositions 6 and 7 make it clear that nothing depends on the fixed error rate ε. Let D_* be the set of data generating processes with the property that, for every outcome h, there exists a pair $(\underline{\rho}, \overline{\rho} \in [0, 1/2) \times (1/2, 1]$, such that

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T_+(h(n))} \sum_{n=1}^{T-1} d_+(h_n) &= \overline{\rho} \\ \lim_{T\to\infty} \frac{1}{T_-(h(n))} \sum_{n=1}^{T-1} d_-(h_n) &= \underline{\rho}, \end{aligned}$$

where $d_+(h_n)$ equals $d(h_n)$ if the latter exceeds $1/2$ and is zero otherwise, $d_-(h_n)$ is analogous for values of $d(h_n)$ less than $1/2$, $T_+(h(n))$ is the number of times theory d has produced a prediction exceeding $1/2$ on the history h_n, and $T_-((n))$ is analogous for predictions less than $1/2$. We are thus assuming that the average error rate in the data generating process, when predicting either 1 or 0, converges (though not necessarily to the same limits). If this is not the case, there is no hope for the agent to identify the appropriate error rates for effective learning. Then arguments analogous to those giving Proposition 7 allow us to establish that for every subjective order $\succcurlyeq^S$, there exists a strategy $\succsim^{S*}$ such that the agent's limiting payoff in periods in which a 1 is predicted approaches $\overline{\rho}$ and the agent's limiting payoff in periods in in which a 0 is predicted approaches $\underline{\rho}$.

4.7 Smooth Trade-Offs

Our central result is that effective learning couples concerns about a theory's likelihood with an auxiliary subjective criterion. Studies of model selection in statistics and in machine learning often similarly suggest a trade-off between likelihood and simplicity. Simplicity takes the place of our subjective order in these criteria, while our lexicographic criterion is typically replaced by a smooth objective function. For example, the Akaike Information Criterion (Akaike [1]) is given by

$$\log(L(t)) - 2k,$$

where $L(t)$ is the likelihood function of theory t and k is the number of parameters used in model t. Related to Kolmogorov's complexity measure (Kolmogorov [7, 8], Chaitin [3], Solomonoff [15]), the Minimal Message Length criterion (Rissanen [10], Wallace and Boulton [19]) suggests

$$\log(L(t)) - MDL(t),$$

where the $MDL(t)$ is the minimum description length of the theory t. (See also Wallace [18] and Wallace and Dowe [20].)

The general form of these measures is

$$\log L(t) - \alpha C(t), \tag{9}$$

where $C(t)$ is a "complexity function" (i.e., a function satisfying the properties prescribed by (4)) and α a constant determining the relative weights placed on the likelihood and on the complexity of the theory. Gilboa and Schmeidler [4] offer an axiomatization of this criterion. In their model the reasoner has an order over theories given data, akin to $\succsim_h$ in our case. Certain axioms on the way theories are ranked by this relation for different histories h imply an additive trade-off between the log-likelihood and a parameter of the theory that may be interpreted as its measure of complexity.

We cannot apply (9), designed to evaluate theories given a fixed set of data, directly to our setting. As we have noted, the likelihood $L(t)$ inevitably declines to zero and hence its log decreases without bound as observations accumulate. This ensures that complexity considerations or any other subjective considerations would eventually play no role in the analysis. We accordingly examine

$$l(t,h) - \alpha C(t), \tag{10}$$

ensuring that likelihood and complexity considerations remain on a common footing.[12]

We can draw a connection between smooth measures such as (10) and our lexicographic criterion. Fix a complexity function $C(t)$ and parameter α, and let $\succsim^{\alpha}$ be the resulting order over theories induced by (10). How does $\succsim^{\alpha}$ compare to $\succsim^{LS}$, where the latter is based on the subjective order over theories induced by the complexity function $C(t)$?

To simplify the discussion, let us restrict attention to a set of data generating processes $D_{\varepsilon}^{C} \subset D_{\varepsilon}$ with the property that for any $d, d' \in D_{\varepsilon}^{C}$, the average log likelihood ratio $l(d', h_n)$ converges with probability one, when the data generating process is d. If we did not do this, $\succsim^{\alpha}$ could fall prey to instability of the type presented in Example 6, and would have to be supplemented by the type of stability criterion presented in Section 4.3 to be effective. Doing so would be straightforward, but would clutter the argument.

Proposition 8 *Let $D \subset D_{\varepsilon}^{C}$ be countable. Then*

$$\lim_{\alpha \to 0} \Pi(d, \succsim^{\alpha}) = 1 - \varepsilon.$$

For a fixed α, the criterion $L(t) - \alpha C(t)$ restricts attention to a finite subset of D_{ε}^{C} as possible maximizers of $L(t) - \alpha C(t)$, since a theory that is too complex can never amass a likelihood value large enough to exceed the value $L(t) - \alpha C(t)$ attained by the simplest theory. Among this finite set, no theory can consistently achieve a likelihood above $1 - \varepsilon$. If α is too large, this finite set will exclude the data generating process itself, and all of the eligible theories may well fall short of likelihood $1 - \varepsilon$. Smaller values of α will not exclude the data generating process *a priori*, but may still lead to the selection of a simpler theory and an attendant likelihood loss. As α gets arbitrarily small, we can be assured that the data generating process is encompassed in the set of eligible theories and that very little likelihood is sacrificed in the interests of simplicity, leading to a payoff approaching $1 - \varepsilon$.

Notice, however, that $\Pi(d, \succsim^{0}) = \Pi(d, \succsim^{L})$, and hence $\Pi(d, \succsim^{0})$ equals $1/2$ (given Assumptions 2 and 4.3). In addition, we cannot say *a priori* how small α must be in order to ensure that $\Pi(d, \succsim^{\alpha})$ is close to $1 - \varepsilon$. We thus need to make α arbitrarily close to zero, without actually equalling zero.

[12] In so doing, we move close to criteria such as the Schwarz Information Criterion (also known as the Bayesian Information Criterion (Schwarz [12])), which retains the additive trade-off but uses a complexity measure that depends on the number of observations.

This is just what our lexicographic criterion does. We can accordingly view the lexicographic criterion as the limiting case of the smooth criteria that have been offered in the literature.

5 Discussion

This section explores several aspects of our model and results. To keep the discussion simple, we present formal results in Sections 5.1–5.3 for the case of a deterministic data generating process.

5.1 Countability

We have assumed the set of data generating processes D is countable. The countability of D may seem quite restrictive. Indeed, most statistical models allow continuous parameters, and thereby seemingly refer to uncountable families of processes. However, our inclination is to be persuaded by Church's thesis—if the agent can make a particular set of predictions, then there must be a Turing machine generating these predictions (Hopcraft and Ullman [6, Chapter 7]), and hence the set of conceivable data generating processes can reasonably be taken to be countable.[13]

But this limitation on the agent's cognitive abilities need not be shared by the set of possible data generating processes. To make this distinction, let D be the set of possible data generating processes, and T the set of theories of which the agent can conceive. We may then have a set D that is an (uncountable) superset of T. How will the agent fare then? Worse still, what if the data generating process is malevolent, using a (noncomputable) strategy that predicts the agent's (computable) predictions in order to then

[13]Alternatively, one may arrive at countability via a more lenient model, in which a Turing machine (or, equivalently, a PASCAL program) can also perform algebraic operations on arbitrary real-valued variables, where the actual computations of these operations are performed by an "oracle" that is not part of the machine's computation. A stricter interpretation of computability, which does not resort to "oracles," would restrict attention to statistical models in which all parameters are computable numbers. A number $x \in \mathbb{R}$ is *computable* if there exists a Turing machine M that, given the description of any rational $\varepsilon > 0$, performs a computation that halts, and writes a number $M(\varepsilon) \in \mathbb{Q}$ such that $|M(\varepsilon) - x| < \varepsilon$. All rational numbers are computable, but so is any irrational number that can be described by a well-defined algorithm, including algebraic irrational numbers (such as $\sqrt{2}$), e, and π.

generate unpredicted observations? To investigate this possibility, we retain the assumption that $T \subset D_0$ is countable, but allow $D \subset D_0$ to be a superset of T.

The standard way for the agent to protect himself against a malevolent data generating process is to randomize. Specifically, for a discriminating subjective order $\succcurlyeq^S$ and for $\varepsilon > 0$, let the relation $\succsim^{LS,\varepsilon}$ be defined by augmenting $\succsim^{LS}$ with a "safety net." If the average payoff at history h_n is lower than $0.5 - \varepsilon/\log n$, then $\succsim^{LS,\varepsilon}_{h_n} = T \times T$. Otherwise, $\succsim^{LS,\varepsilon}_{h_n} = \succsim^{LS}_{h_n}$.

Proposition 9 *Let $T \subset D_0$ be countable. Let Assumption 2 hold and let T satisfy Assumptions 1.1 and 1.3 (while allowing $D \subset D_0$ to be a superset of T). Then $\succsim^{LS,\varepsilon}$ weakly dominates $\succsim^{L}$ for every discriminating subjective relation $\succcurlyeq^S$, with $\succsim^{LS,\varepsilon}$ performing strictly better for data generating processes $d \in T$.*

We can think of the relation $\succsim^{LS,\varepsilon}$ as mimicking the relation $\succsim^{LS}$ as long as "all goes well." All will go well, and the use of the discriminating subjective order $\succcurlyeq^S$ will then ensure a payoff approaching unity, whenever the data generating process is drawn from T. This will also be the case for many data generating processes drawn from outside the set T. The signal that things are not going well is an average payoff that dips below 1/2. In this event, the agent resorts to randomizing equally over predicting 0 and predicting 1. This ensures a payoff of 1/2, which is at least as large as that garnered by the likelihood relation. The agent thus effectively learns when learning is possible, and otherwise does not do worse than if no subjective considerations entered the prediction process.

5.2 Computability

We have justified the assumption that the set of conceivable theories is countable by appealing to computability arguments, in the form of an assumption that the agent can only implement predictions generated by a Turing machine. Continuing in this spirit, we now take computability issues more seriously. Let us first restrict the data generating process to the set D_0^H of deterministic data generating processes implementable by Turing machines that halt after every input $h \in H$.

In contrast, we now allow the agent to consider the set D_0^T of all Turing machines, even those that do not always halt. It is a relatively easy task for

the agent to enumerate all Turing machines, but it is not an easy task to check which of them do indeed define a data generating process.[14] A model that respects the agents' computability constraints must then allow the set T to include *pseudo-theories*: all machines that can be written in a certain language (and therefore appear to define a data generating process), even if they may not halt for all histories. Clearly, this additional freedom cannot help the agent: if, at a given history h, the agent chooses a machine that does not halt for that history, she will never be able to make a prediction (in which case we take her payoff to be 0). However, "helping" the agent by assuming that $T \subset D_0^H$ would be unreasonable, as it would be tantamount to magically endowing the agent with the ability to solve the celebrated halting problem.[15]

We also restrict the agent to relations $\succsim$ that are computable, in the sense that for every $h \in H$, the choice made by the relation $\succsim_h$ from the set $B_{\succsim_h} \subset D_0^T$ could itself be implemented by a Turing machine that inevitably halts. This restriction is a binding constraint for some data generating processes:

Proposition 10 *For every computable relation $\succsim \subset D_0^T \times D_0^T$, there exists a data generating process $d \in D_0^H$ such that $\Pi(d, \succsim) \leq 0.5$.*

Proposition 10 imposes a bound on what can be guaranteed by a computable strategy, in the sense that any such strategy must fare no better than chance against *some* data generating processes. The proof consists of observing that if the agent's strategy is computable, then one may always construct a malevolent strategy d that mimics the agent's computation and chooses an observation that refutes it.

The malevolent strategy d used to prove Proposition 10 is quite far from most statistical models. In particular, it is counterintuitive to imagine the world simulating the agent's reasoning, not to mention refuting the resulting

[14] One could simulate the computation of any given machine given input h, but there is no way to distinguish between computations that take a long time and computations that never end.

[15] Formally speaking, the objects of choice for the agent are not theories but descriptions thereof. A rigorous treatment of this problem would call for the definition of a formal language and of a means of describing programs in that language. Some descriptions give rise to well-defined theories (i.e., that halt for every history), whereas others would not. In such a model, every theory would have infinitely many equivalent descriptions. Thus, the function that maps descriptions to theories is not defined for all descriptions and is not one-to-one.

belief period after period. Will a more neutral model of the data generating process allow a possibility result? One way to obtain a more realistic set of data generating processes is to limit their computations. Specifically, let D_0^B be the set of data generating processes that are implementable by Turing machines that halt within a bounded number of steps. That is, for $d \in D_0^B$ there exists a Turing machine $M(d)$ and an integer $K(d)$ such that, for every history h_n and attendant prediction y_n, the computation of $M(d)$ on h_n halts within $K(d)$ steps, producing y_n.

The agent is restricted to have a (discriminating) subjective order that is represented by a computable function $C : D_0^T \to \mathbb{N}$, so that

$$C(t) \leq C(t') \qquad \Longleftrightarrow \qquad t \succcurlyeq^S t'.$$

Thus, because C is computable, the agent can compute $\succcurlyeq^S$.

The following result adapts subjective-based rankings to the computable set-up.

Proposition 11 *For every computable subjective order $\succcurlyeq^S \subset D_0^T \times D_0^T$, there exists a computable relation $\succsim$ with each $\succsim_h \subset D_0^T \times D_0^T$ such that*

(11.1) $\Pi(d, \succsim) = 1$ for every $d \in D_0^B$;

(11.2) for every $d, d' \in D_0^B$ there exists N such that, for every $n \geq N$ and every $h \in H_n$ for which $L(d, h) = L(d', h)$,

$$d \succ_h^S d' \Rightarrow d \succ_h d'.$$

Proposition 11 ensures the existence of a computable strategy yielding optimal payoffs, as well as its asymptotic agreement with the (strict part of) the given subjective ordering $\succcurlyeq^S$ over D_0^T.[16] The relation $\succsim$ cannot follow $\succsim^{LS}$ precisely, but it does so for long enough histories. In other words, it is possible that for a short history the relation $\succsim$ will not reflect the subjective ranking $\succcurlyeq^S$, but in the long run, any two theories that are equally accurate but not equally simple will be ranked according to $\succcurlyeq^S$.

Observe that most deterministic statistical models encountered in the social sciences are in D_0^B. The deterministic version of models such as linear

[16]In a context focusing on computability, it would be natural to think of $d \succcurlyeq^S d'$ if the Kolmogorov complexity of d is lower than that of d', i.e., if d has a shorter minimal description length than d'. This still leaves some freedom in defining $\succcurlyeq^S$. For instance, one may choose a description in a given programming language, such as PASCAL, as opposed to Turing machines, and one may take the description of constant values into account in the measurement of the description length, or decide to ignore them, and so on.

regression, non-linear regression, as well as many models in machine learning, can be described by an algorithmic rule whose computation time does not depend on the input. A notable exception are time series in economics, where the model describes the dependence of y_n on $\{y_i\}_{i<n}$, and thus the length of the computation increases with the length of history, n.

5.3 Impatience

Suppose that the agent has a discounted payoff criterion,

$$\Pi^\delta(d, \succsim) = \mathcal{E}\left\{(1-\delta)\sum_{n=0}^{\infty}\delta^n \pi(d, \hat{t}_n, h_n)\right\}. \tag{11}$$

We assume the data generating process is chosen according to a probability measure λ on D.

It is advantageous to use the subjective order provided that the agent is sufficiently patient. In particular, the (omitted) proof of the following is a straightforward modification of the arguments used to prove Propositions 1 and 2:

Proposition 12 *Let Assumptions 1 and 2 hold, let payoffs be given by (11), and let $\succcurlyeq^S$ be a discriminating subjective order.*

[12.1] For every $d \in D$, there is a discount factor δ^* such that for all $\delta \geq \delta*$,

$$\Pi^\delta(d, \succsim^{LS}) > P^\delta(d, \succsim^{L}).$$

[12.2] If the data generating process is chosen according to density λ on D, then there is δ^* such that for all $\delta > \delta^*$,

$$\int_D \Pi^\delta(d, \succsim^{LS})d\lambda > \int_D \Pi^\delta(d, \succsim^{L})d\lambda.$$

5.4 Simplicity

There are many sources of subjective biases that distinguish among theories. We are especially intrigued by the possibility that a preference for simplicity may play a role in subjective evaluations. A preference for simplicity is

among the most universal criteria for theory selection—people tend to prefer simpler explanations and simpler theories to more complex ones.[17]

The notion of simplicity raises several fundamental questions: What does it mean to say that "theory t is simpler than theory t'?" To what extent can such an ordering be viewed as objective? Where does the ordering come from in the first place? These questions have been the subject of an immense literature (e.g., Sober [14]). Importantly, ever since Goodman [5] presented the "grue-bleen" paradox, it has been evident that the notion of simplicity is language-dependent and that it eludes obvious definition. Indeed, Kolmogorov's operationalization of the notion of simplicity (Kolmogorov [7], Chaitin [3]) clarifies that using different languages as primitives can lead to different simplicity orderings (though there are some limitations on the divergence of different orderings if the relevant languages can be translated to each other (see Solomonoff [15])). We hope that the framework provided in this paper may be used to investigate the role of simplicity in inductive inference.

6 Appendix: Proofs

Proof of Proposition 1 Assumption 1.3 ensures that, for every history h_n there are theories $t \in D$ consistent with h_n, that is, theories satisfying $L(t, h_n) = 1$. Consider the set of such theories,

$$B_{\succsim^L_h} = \{\, d \,|\, L(d, h) = 1 \,\}.$$

For any finite continuation of h_n there is a theory $t \in B_{\succsim^L_h}$ that is consistent with this continuation. In particular, this is true for the history h_{n+1} generated from h_n and $y_n = 0$ (coupled with x_{n+1}) as well as for the history h'_{n+1} generated from h_n and $y_n = 1$ (coupled with x_{n+1}). Assumption 2 then ensures that the order $\succsim^L$ is equally likely to select a theory predicting $y_n = 0$ as it is to select a theory $y_n = 1$. Thus, the probability of making a correct prediction is $1/2$, and hence $\pi(d, h_n, t_n) = 0.5$, regardless of the true process d. This establishes

$$\Pi(d, \succsim^L_h) = 0.5.$$

■

[17] This preference for simplicity has been championed on normative grounds (most famously by William of Occam (see Russell [11])) and has long been offered as a descriptive model of human reasoning (e.g., Wittgenstein [21]).

Proof of Proposition 2 Fix $d \in D$. Since $\succcurlyeq^S$ is discriminating, there are finitely many theories in $S(d) \equiv \left\{t \in D \,|\, t \succcurlyeq^S d\right\}$, i.e., that are ranked ahead of or indifferent to d by the subjective order $\succcurlyeq^S$. Choose some $t \in S(d)$ and suppose that t and d are not observationally equivalent, meaning that they do not generate identical outcomes $((x_0, y_0), ..., (x_n, y_n), ...)$. Then at some period n theory t will be refuted, i.e., the data generating process will produce a history $h_n = ((x_0, y_0), ..., (x_{n-1}, y_{n-1}), x_n)$ for which $L(t, h_n) = 0$ and hence $d \succ_h^{LS} t$. Applying this argument to the finitely many theories in $S(d)$, there must exist a finite time n' by which either theory d is chosen by $\succsim_{h_{n'}}^{LS}$ or some element $t \in S(d)$ is chosen by $\succsim_{h_{n'}}^{LS}$ that is observationally equivalent to d. Thereafter, $\pi(d, t_n, h_n) = 1$ holds. This yields $\Pi(d, \succsim^{LS}) = 1$. We conclude that, for every relation $\succsim^S$ derived from a discriminating subjective order $\succcurlyeq^S$, and for every data generating process d, the limit payoff under $\succsim^{LS}$ is 1, while it is only 0.5 under $\succsim^L$. Hence, $\succsim^{LS}$ strictly dominates $\succsim^L$. ■

Proof of Proposition 3. Consider an agent characterized by $\succsim^{LI}$ and a data generating process d. If $\Pi(d, \succsim^{LI}) < 1$, it must be that infinitely often, $\pi(d, t_j, h_j) = 0$. Hence, the agent infinitely often chooses a new theory but never chooses d. By Assumption 3, the probability that the agent chooses a new theory n times without choosing d is at most $(1 - \lambda(d))^n$. Since $\lim_{n\to\infty}(1-\lambda(d))^n = 0$, the probability that $\Pi(d, \succsim^{LI}) < 1$ is zero, and hence the expected value of $\Pi(d, \succsim^{LI})$ is unity. ■

Proof of Proposition 6. Fix a data generating process d. Assume that γ satisfies $\theta(\gamma) = \theta(1 - \varepsilon) - \delta$ for $\delta > 0$. For any $\eta > 0$, there exists k such that, with probability $1 - \eta$ at least, for all $n \geq k$,

$$l(d, h_n) > [(1 - \varepsilon)\log(1 - \varepsilon) + \varepsilon \log \varepsilon] - \delta = \theta(1 - \varepsilon) - \delta = \theta(\gamma).$$

Thus, from period k on, it is likely that the correct theory d is among the γ-maximizers of $l(\cdot, h_n)$. If d is the maximizer of $\succsim^{LS,\gamma k}$ used for prediction, a payoff of $(1 - \varepsilon)$ is guaranteed. We wish to show that, if another theory is used for prediction, it cannot be much worse than d itself.

Let us condition on the probability $1 - \eta$ event that for every $n > k$, $l(d, h_n) > \theta(\gamma)$. If a theory $t \neq d$ is used for prediction at period $n \geq k$, then it must be the case that (i) t is a γ-best fit for all periods $j = k, ..., n$; and (ii) $t \succcurlyeq^S d$. Hence, for each period $n > k$, there are only a finite number of theories satisfying conditions (i) and (ii), of which the highest-ranked

by the subjective order will be chosen. Moreover, the set of such theories is decreasing in n (since a theory whose likelihood ratio drops below γ is subsequently disqualified). Eventually, a period n' will be reached such that some theory t (possibly d) satisfying (i) and (ii) will be used in all subsequent periods. Let $n > n'$, and let α be the proportion of times, up to n, that t made the correct prediction. Then, since t is a γ-best fit at n, we have

$$\begin{aligned} l(t,h) &= \alpha \log(1-\varepsilon) + (1-\alpha)\log\varepsilon \\ &= \alpha\left[\log(1-\varepsilon) - \log\varepsilon\right] + \log\varepsilon \\ &= \alpha \log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon \\ &\geq \theta(\gamma) \\ &= \theta(1-\varepsilon) - \delta \\ &= (1-\varepsilon)\log(1-\varepsilon) + \varepsilon\log\varepsilon - \delta \\ &= (1-\varepsilon)\left[\log(1-\varepsilon) - \log\varepsilon\right] + \log\varepsilon - \delta \\ &= (1-\varepsilon)\log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon - \delta. \end{aligned}$$

This gives

$$\alpha \log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon \geq (1-\varepsilon)\log\frac{1-\varepsilon}{\varepsilon} + \log\varepsilon - \delta$$

or

$$[\alpha - (1-\varepsilon)]\log\frac{1-\varepsilon}{\varepsilon} \geq -\delta$$

that is,

$$\alpha \geq (1-\varepsilon) - \frac{\delta}{\log\frac{1-\varepsilon}{\varepsilon}}.$$

Intuitively, the payoff obtained by predicting according to t cannot be much lower than $(1-\varepsilon)$. Taking into account the probability of convergence by time k we get

$$\Pi(d, \succsim^{LS,\gamma k}) \geq (1-\eta)\left[(1-\varepsilon) - \frac{\delta}{\log\frac{1-\varepsilon}{\varepsilon}}\right],$$

which converges to $(1-\eta)(1-\varepsilon)$ as $\delta \searrow 0$. Finally, increasing k results in decreasing η to any desired degree, and the result follows. ■

Proof of Proposition 7. The basic idea is have the agent simulate the choices of theories that would have corresponded to $\succsim^{LS,\gamma k}$ for different values of γ and of k. For values of γ larger than $1-\varepsilon$, the agent will find that the maximizers of $\succsim^{LS,\gamma k}$ keep changing, indicating that γ is too high. For values of γ that are lower than $1-\varepsilon$, the agent will find theories that get selected asymptotically, an indication that γ might be too low. By refining the search for γ, while simultaneously gathering more observations, the reasoner will approach $1-\varepsilon$ and make predictions according to the correct theory.

We make these ideas precise in the form of a reasoning algorithm that is simple, but makes no claims to efficiency. At stage n the reasoner considers as possibilities for γ all values in

$$\Gamma_n = \left\{ \frac{r}{2^n} \,|\, r = 0, 1, ..., 2^n \right\}.$$

Given n, define $k = \llcorner n/2 \lrcorner$. For each $\gamma \in \Gamma_n$, and for each $m = k, ..., n$, the reasoner finds all the maximizers of $\succsim_{h_m}^{LS,\gamma k}$ (to make this an algorithm, we need to assume that an oracle can perform this task). Denote the set of maximizers for each γ by $M(m, k, \gamma)$. This is a finite set, due to the agent's preference for simplicity. Then, for each γ, define

$$M^*(n, \gamma) = \cap_{k \leq m \leq n} M(m, k, \gamma).$$

Thus, $M^*(n, \gamma)$ contains precisely those theories that have been among the "γ-best" theories for the past $n/2$ periods.

If $M^*(n, \gamma) = \varnothing$ for all $\gamma \in \Gamma_n$, define $\succsim_{h_n}^{S*} = D \times D$. In this case all theories are equivalent in terms of $\succsim_{h_n}^{S*}$, and the reasoner's choice will be arbitrary.

If, however, $M^*(n, \gamma) \neq \varnothing$ for some $\gamma \in \Gamma_n$, let γ_n be the maximal such value in Γ_n, and define

$$t \succsim_{h_n}^{S*} t' \quad \Longleftrightarrow \quad \left\{ \begin{array}{rrcll} & t & \in & M^*(n, \gamma_n) & \text{and} \quad t' \notin M^*(n, \gamma_n) \\ \text{or} & t, t' & \in & M^*(n, \gamma_n) & \text{and} \quad t \succcurlyeq^S t' \\ \text{or} & t, t' & \notin & M^*(n, \gamma_n). & \end{array} \right.$$

That is, the $\succcurlyeq^S$-most-prefered theories in $M^*(n, \gamma_n)$ are considered to be the "best" theories and one of them will be used for prediction.

To see that the definition of $\succsim^{S*}$ satisfies the desired properties, observe that, by the proof of Proposition 6, if $\gamma > 1 - \varepsilon$, $M^*(n, \gamma) = \varnothing$ for large n.

For $\gamma < 1-\varepsilon$, $d \in M^*(n,\gamma)$ for large n. As $n \to \infty$, the minimal γ for which $M^*(n,\gamma) \neq \varnothing$ converges to $1-\varepsilon$, and d is among the maximizers of $\succsim^{S*}$. We then repeat the argument of Proposition 6, by which any theory $t \neq d$ such that $t \in M^*(n,\gamma)$ obtains a payoff that converges to $(1-\varepsilon)$ as $\gamma \nearrow 1-\varepsilon$. ■

Proof of Proposition 8. Fix a complexity function $C(t)$, a value $\alpha > 0$, and a data generating process d^*. Let $\hat{d} \in \arg\min_{d \in D^C_\varepsilon} C(d)$. Then no theory d for which $\theta(1) - \alpha C(d) < \theta(\varepsilon) - \alpha C(\hat{d})$ will ever be chosen by the relation $\succsim^\alpha$, no matter what the history. The agent's choice of theory in each period will thus be drawn from the finite set $D^C_\varepsilon(\alpha) \equiv \{d \in D^C_\varepsilon : \theta(1) - \alpha C(d) < \theta(\varepsilon) - \alpha C(\hat{d})\}$.

For sufficiently small α, the data generating process d^* is contained in $D^C_\varepsilon(\alpha)$. In addition, with probability 1, the limit $\lim_{n\to\infty} l(d,h_n)$ exists for all $d \in D^C_\varepsilon(\alpha)$. Since this set is finite, with probability 1, the agent's choice of theory becomes constant across periods, being the maximizer over $D^C_\varepsilon(\alpha)$ of

$$\lim_{n\to\infty} l(d,h_n) - \alpha C(d).$$

But since $d^* \in D^C_\varepsilon(\alpha)$ for small α, the agent's payoff is bounded below by

$$\lim_{n\to\infty} l(d,h_n) - \alpha C(d) = \theta(1-\varepsilon) - \alpha C(d^*).$$

Taking α to zero then gives the result. ■

Proof of Proposition 9. The relation $T \times T$ guarantees a random choice (by Assumption 1.3), and hence this relation ensures an expected payoff of 0.5 at each period in which it is played. Thus, if $\succsim^{LS,\varepsilon} = T \times T$ for a long enough period, the average payoff converges to 0.5 with probability 1. Moreover, it does so at a rate proportional to $n^{-1/2}$. It follows that, with probability 1, the sustained application of relation $T \times T$ leads to a period n at which the average payoff surpasses the threshold $0.5 - \varepsilon/\log n$, at which point $\succsim^{LS,\varepsilon}$ switches to $\succsim^{LS}$.

Suppose $d \in T$. Since $\succsim^{LS,\varepsilon} = \succsim^{LS}$ infinitely often, $\succsim^{LS,\varepsilon}$ will eventually select d or a theory equivalent to d. Predictions will subsequently be perfect, ensuring that $\succsim^{LS,\varepsilon}$ will not revert to $T \times T$ and that $\Pi(d, \succsim^{LS,\varepsilon}) = 1$.

If $d \notin T$, the lowest the average payoff at history h_n can drop without ensuring $\succsim^{LS,\varepsilon} = T \times T$ is $0.5 - \varepsilon/\log n - 1/n$ (obtained by coupling a history of length $n-1$ in which the payoff is precisely $0.5 - \varepsilon/\log(n-1)$ with one

more incorrect observation). Hence $\Pi(d, \succsim^{LS,\varepsilon}) \geq 0.5$. Combining the two, we thus find that $\Pi(d, \succsim^{LS,\varepsilon}) \geq \Pi(d, \succsim^{L})$ for all $d \in D$, with strict equality for every $d \in T$. ■

Proof of Proposition 10. Let $\succsim$ be computable. Then there is a Turing machine τ that implements $\succsim$ by, for any history h, computing a maximizer of $\succsim$ from the set D^H. Let d simulate the machine τ, for any history h, finding the maximizer t_h that the agent will use for prediction, and then generating prediction 1 if $t_h(h) \leq 0.5$ and 0 if $t_h(h) > 0.5$. A deterministic t will result in a payoff of 0. The maximal payoff for the agent at each period is 0.5, obtained by the random prediction $t_h(h) = 0.5$. ■

Proof of Proposition 11. The basic idea is to construct the relation $\succsim$ by combining the underlying subjective order $\succcurlyeq^S$ with the time complexity of the machine.

Let $D_0^T = \{t_1, t_2, ...\}$ be the class of all Turing machines, including those that always halt and those that do not halt for certain inputs $h \in H$. There is no difficulty in writing a machine that generates D_0^T, or, equivalently, that can accept $i \geq 1$ as an input and, after a finite number of steps, provide the description of t_i.

Assume we are given a history h and we wish to select a theory that has high likelihood and that halts for h. When considering a machine t, we thus need to determine whether it fits the data, namely whether $L(t, h_n) = 1$ (taking $L(t, h_n) = 0$ if the machine fails to halt for any prefix of h_n), and we need to compute its prediction for y_n, or $t(h_n)$, taking into account the possibility that it may not halt when making this prediction. That is, we need to know the result of $n + 1$ computations of t_i (one to verify that the theory fits the observation generated in each of the preceding n periods, and one to generate the current prediction), each of which may not halt.

Let $C : D \to \mathbb{N}$ be a computable complexity function for the underlying subjective order $\succcurlyeq^S$, so that

$$C(t) \leq C(t') \qquad \Longleftrightarrow \qquad t \succcurlyeq^S t'.$$

Define $c : D \times H \to \mathbb{N} \cup \{\infty\}$ to be the length of computation, that is, $c(t, h) \in \{1, 2, ..., \infty\}$ is the number of steps that t takes to compute where

h is its input. Next define a function $C^* : D \times H \to \mathbb{R}_+ \cup \{\infty\}$ by

$$C^*(t,h) = C(t) + \frac{1}{n^2}\sum_{j=0}^{n} c(t,h_j)$$

where $t \in D$, $h \in H_n$ and h_j is the j-th prefix of h. Using this function, we define our candidate relation over theories:

$$t' \succsim_h t \qquad \Longleftrightarrow \qquad \begin{cases} & L(t',h) > L(t,h) \\ or & [L(t',h) = L(t,h) \;\; and \;\; C^*(t',h) \leq C^*(t,h)] \end{cases}.$$

We argue that it is a computable task to find a maximizer of $\succsim_h$ from among those machines that halt on history h, and that this maximizer will have likelihood one. First observe that for every h there exists a machine t such that $L(t,h_n) = 1$ and $C^*(t,h_n) < \infty$. To see this, it suffices to consider a machine t that generates history h_n irrespective of the data. For any history longer than n, the machine can generate 0. This takes a computation time $c(t,h) = O(n)$. By construction, $t \in D_0^B$. Since this machine appears somewhere in the enumeration corresponding to $\succ^S$, we have $C(t) < \infty$ and hence $C^*(t,h) < \infty$.

Given $C^*(t,h)$, there are finitely many machines t' with $C(t') \leq C^*(t,h)$, and therefore only finitely many machines that can beat t according to $\succsim$. Each of these has to be simulated only a bounded number of steps, $C^*(t,h)$, to see if, indeed, it gives $L(t',h_n) = 1$ and a lower value for $C^*(t',h)$.

Note that for all $d \in D_0^B$, $c(t,h_n) \leq K(d)$ and

$$\frac{1}{n^2}\sum_{j=0}^{n} c(t,h_j) \leq \frac{1}{n^2} nK(d) \to 0$$

hence,

$$C^*(t,h) \to C(t).$$

Now consider $d, d' \in D_0^B$ with $d \succ^S d'$ and hence $C(d) \leq C(d')$. Then for all sufficiently large n, $C^*(d,h_n) < C^*(d',h_n)$, and hence $L(d,h_n) = L(d',h_n) \Rightarrow d \succ_h d'$. This establishes (11.2).

We now turn to (11.1), namely that $\Pi(\succsim, d) = 1$ for every $d \in D_0^B$. For $t' \succsim_h d$ to hold, we must have $L(t',h) = 1$ and $C(t') \leq C(d)$. An argument analogous to that of the proof of Proposition 2 ensures that at some point, d or a theory equivalent to it is found, and from that point on only such theories (predicting $d(h)$ for every h) can be maximizers of $\succsim$. Hence the agent makes perfect predictions and obtains $\Pi(\succsim, d) = 1$. ■

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Nabil Ibraheem Al-Najjar. Decision makers as statisticians. Technical report, Northwestern University, 2008.

[3] Gregory J. Chaitin. On the length of programs for computing binary sequences. *Journal of the Association for Computing Machinery*, 13(4):547–569, 1966.

[4] Itzhak Gilboa and David Schmeidler. Likelihood and simplicity: An axiomatic approach. Mimeo, Tel Aviv University, 2008.

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

[6] John E. Hopcraft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison Wesley, Reading, Mass., 1979.

[7] Andrei. N. Kolmogorov. Three approaches to the quantitative definition of information. *Probability and Information Transmission*, 2(1):4–7, 1965.

[8] Andrei. N. Kolmogorov. On tables of random numbers. *Theoretical Computer Science*, 207(6):387–395, 1998 (originally 1963).

[9] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1996.

[10] Jorma Rissanen. Modelling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[11] Bertrand Russell. *History of Western Philosophy*. Routledge, London, 2004. Originally 1946.

[12] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[13] Herbert Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69:99–118, 1955.

[14] Elliott Sober. *Simplicity*. Clarendon Press, Oxford, 1975.

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

[16] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[17] Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[18] Christopher S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, New York, 2005.

[19] Christopher S. Wallace and D. M. Boulton. An information measure of classification. *The Computer Journal*, 13:185–194, 1968.

[20] Christopher S. Wallace and David L. Dowe. Minimal message length and Kolmogorov complexity. *The Computer Journal*, 42:270–283, 1999.

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