Subjectivity in inductive inference

Itzhak Gilboa
HEC, Paris and School of Economics, Tel Aviv University

Larry Samuelson
Department of Economics, Yale University

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.

Keywords. Inductive inference, simplicity, prediction, learning.

JEL classification. C1, D8.

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 have seen, which rule (or "theory" or "model") should I use to predict future observations?

A first fundamental principle is that one should consider only 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, or 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? Admittedly, perfectly objective inference is impossible: even the purest of classical statisticians

Itzhak Gilboa: tzachigilboa@gmail.com

Larry Samuelson: Larry.Samuelson@yale.edu

must exercise judgement when deciding which variables to include in a model, which significance level to choose, and so forth. But should we not 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 objective arguments alone: simply observing that one theory fits the data better than another is not sufficient to prefer the former to the latter. Instead, one must also argue that the candidate theory fares well in terms of consistently applied subjective auxiliary criteria.

1.2 *Objectivity and subjectivity*

We consider an agent who faces a sequence of prediction problems. In each period, she observes an input and must predict the value of an output. To make her prediction, the agent selects a theory, which specifies a distribution over possible predictions, given the history of past inputs and outputs.

The agent's choice of a theory in each period is made according to a preference relation over theories. We are interested in the degree of objectivity of such criteria, as defined by the degree to which they depend on data, that is, on the history of past inputs and outputs. We say that the agent's preference over theories is *(purely) objective* if it depends only on the conditional probabilities of past observations given theories. For example, if theories are ranked according to the likelihood function, the agent's preferences are objective. However, one may choose to maximize other functions of the conditional probabilities, such as the minimal probability assigned by a theory to the observations encountered, and the resulting preference is also objective according to our definition.

We say that the agent's preference over theories is *(purely) subjective* if it is independent of the data. Subjective criteria for ranking theories include a preference for simplicity and a preference for intuitive explanations. A Bayesian approach that assigns prior probabilities to theories also defines a subjective criterion. When a Bayesian prior is updated in response to accumulating observations, it defines a criterion that is partly subjective and partly objective.

One might think of other distinctions between objective and subjective. Our notion of objectivity does not specify which function is applied to conditional probabilities, and it thus leaves room for the agent's subjective discretion. Alternatively, one may attempt to find some objectivity in history-independent criteria, such as in agreement on the concept of simplicity, or optimality of a prior belief. However, for our purposes, objectivity is defined by dependence on conditional probabilities alone, and subjectivity is defined by independence of history.

To consider an example, suppose that the agent observes the sequence $1, 2, 4, 8$ and is asked to predict the next observation. Clearly, the deterministic theory that predicts the observation $f(n) = 2^{n-1}$ at period n fits the data perfectly well and predicts that the number 16 comes up next. However, the theory

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

also provides a perfect match to the first four observations, but predicts the number 7. We say that, based on objective criteria alone, one cannot distinguish between f and g. If one prefers f to g based on the relative simplicity of the former, we refer to this as applying a subjective criterion, because the judgement that f is simpler than g can be made before any data are observed and does not change with additional observations. Putting these criteria together, the choice of the simplest theory that matches the data is considered to be a combination of a subjective and an objective criterion.

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 necessarily is effective if and only if it discriminates between theories sufficiently finely, in a sense we make precise below. 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 is effective if goodness-of-fit and subjective considerations are balanced so as to produce some stability in the theories used to predict future observations.

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 alone. This result rests on a simple enumeration argument: the subjective reasoner eliminates incorrect theories until she gets to the correct one, thereafter predicting the environment with precision. In contrast, an agent who seeks only to maximize likelihood may 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 so as 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 thus embraces a theory that promises enhanced explanatory power only if it is sufficiently subjectively appealing *and* has provided a sufficiently good fit for a 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 (i.i.d.) manner. Instead, our main concern 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 prediction of the weather on a given day, assuming that it follows an i.i.d. distribution, our main concern is to determine whether global warming is underway. We are thus interested in a learning problem that is nontrivial even for deterministic processes.

2. The model

We consider a repeated prediction problem. In each period $0, 1, \ldots$, an agent is called on 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.[2]

The observations are produced by a data-generating process, which is 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, and which she then uses to make her prediction. This paper studies the basic problem facing the agent, which is to choose the preference relation over theories that induces her to make payoff-maximizing predictions.

2.1 *Observations*

At the beginning of each period $n \in \{0, 1, \ldots\}$, the agent observes a profile of variables $x_n = (x_n^1, \ldots, x_n^m) \in \{0, 1\}^m \equiv X$. 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

---

[1] For example, the Vapnik–Chervonenkis theory (Vapnik 1998, Vapnik and Chervonenkis 1971), recently applied to decision theory by Al-Najjar (2009), deals with the rate at which one can simultaneously learn the probabilities of multiple events.

[2] An obvious alternative is to consider the discounted sum of expected payoffs. We can obtain approximate versions of our results under this payoff criterion, provided that the agent is sufficiently patient.

D	Set of data-generating processes ($\subset [0,1]^H$).
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 discussion relative to this sequence, without specifying the process that generated it.[3] 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 $h_n = ((x_0, y_0), \ldots, (x_{n-1}, y_{n-1}), x_n)$. 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 $H = \bigcup_{n\geq0} H_n$, with h denoting an element of H.

2.2 *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 are often interested in problems in which the set of possible data-generating processes is a strict subset of $[0,1]^H$. For example, we initially consider the set $\{d \in [0,1]^H \mid d(h) \in \{0,1\} \ \forall h \in H\} \equiv D_0$ of all deterministic data-generating processes.

In the course of our discussion, we consider several possibilities for the set D. In particular, we are especially interested in cases in which D is a subset of the set of data-generating processes considered possible by the agent, or (equivalently, in our view) that the agent is capable of conceiving or understanding. We do this not because we are convinced the agent can understand anything that might occur in Nature, but because learning will be a hopeless task if the data-generating process lies outside the set considered possible by the agent. Since our argument is that objective criteria may not suffice for learning, we focus on a setting in which learning is possible. It is useful for future reference to collect the notation for the various sets of data-generating processes in Figure 1.

2.3 *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

---

[3]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.

are obviously inapplicable, while suggesting that others are relatively likely to generate correct predictions.

The basic characteristic of an agent is a collection of relations $\{\succsim_h \subset D \times D \mid h \in H\} \equiv \succsim$ 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.[4] 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

$$B_{\succsim_h} = \{t \in D \mid t \succsim_h t' \ \forall t' \in D\}$$

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 $\succsim$. Which specifications of $\succsim$ allow the agent to earn high payoffs?

2.4 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

$$\pi(d, t_{h_n}, h_n) = d(h_n)\mathbf{1}_{t_{h_n}(h_n)>0.5} + \big(1 - d(h_n)\big)\mathbf{1}_{t_{h_n}(h_n)<0.5} + \tfrac{1}{2}\mathbf{1}_{t_{h_n}(h_n)=0.5}.$$

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

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

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 an 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 $\succsim$ to choose theories, be given by the Banach limit of the resulting sequence of expected values from (1). The key property of Banach limits we need is

---

[4]Assuming that the set of data-generating processes D is also the set of theories to which the agent appeals when making predictions gives rise to a relatively favorable environment for prediction. It rules out cases in which the agent cannot conceive of some possible data-generating processes (e.g., the agent chooses from a subset of D), in which case effective prediction might well be impossible a priori. Since most of our results are obtained for each possible data-generating process in D, no loss of generality is implied by assuming that this set is at least as large as the set considered by the agent. 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.

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 suffices for our results.

2.5 *The likelihood relation*

The most common objective criterion for theory selection is the *likelihood relation*, ranking theories according to their fit to the data.[5] Formally, define the likelihood of theory t given history h_n:

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

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

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

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 to be simple enough to be learned; the final part requires it to 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.

(i) $D \subset D_0$, the set of deterministic data-generating processes.

---

[5] We focus throughout on the likelihood relation, but analogous results and proofs hold for any objective relation. In particular, no matter what objective relation is used to cull theories, a sufficiently rich set of data-generating processes includes some survivors predicting a 0 as the next observation and other survivors predicting a 1. At this point, the objective relation has exhausted its implications, dooming the agent to random choice.

(ii) D is countable.

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

Given Assumption 1(i), the remaining requirements are satisfied if (for example) D is the set D_0^H of all Turing machines that generate 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 is 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 $\succsim^L$ as a guide to making predictions, but applies no other criteria.

First, note that the agent's choice of theory when using an arbitrary relation $\succsim$ 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 if D is the set D_0^T (of all Turing machines that generate functions $d \in D_0$) or D_0^H (of all Turing machines that do so and halt), then for any history of observations h and for every data-generating process d in D consistent with h, there is another data-generating process in D that is also consistent with h, but whose subsequent datum is precisely the opposite of d, generating a 0 whenever d produces a 1 and vice versa. As a result, the sets $\{t \in B_{\succsim_h} \mid t(h) = 0\}$ and $\{t \in B_{\succsim_h} \mid t(h) = 1\}$ not only are nonempty, but are symmetric in their treatment of the next observation. The obvious implementation of our unbiased-choice provision is then described as follows.[6]

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

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

whenever

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

---

[6]A less demanding version of Assumption 2 assumes only that $\mu_{B_{\succsim_h}}(\{t \in B_{\succsim_h} \mid t(h) < 0.5\})$ and $\mu_{B_{\succsim_h}}(\{t \in B_{\succsim_h} \mid t(h) > 0.5\})$ are both positive when both sets are nonempty. Assumption 2 is used in Propositions 1, 4, 5, and 9. In each case, the role of Assumption 2 is to place a bound on the performance of the likelihood relation. The virtue of Assumption 2 is that it makes this bound particularly simple to state and establish. We could establish analogous but more cumbersome results with the weaker version. The likelihood ratio might then generate a payoff greater than $1/2$ for some data-generating processes, but only at the cost of generating smaller payoffs for other data-generating processes, and the subjective relation would continue to dominate the likelihood relation.

We then have the following proposition.

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

The proof, contained in the Appendix, 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 brings subjective criteria to bear in choosing between theories. To define such a theory-selection procedure, we begin with a subjective order $\succ^S \subset D \times D$.[7] Observe that it is defined a priori, independently of history. We require $\succ^S$ to be complete and transitive. In addition, we say that $\succ^S$ is a *discriminating* subjective order if

$$\#\{t' \in D \mid t' \succ^S t\} < \infty \quad \forall t \in D. \tag{2}$$

Condition (2) 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 is stuck choosing among too many theories that fit the data. The subjective order helps select between such indifferent theories, but is effective only if it does a good enough job of breaking indifferences. In the absence of condition (2), for example, the definition of a subjective order is consistent with the trivial order $\succ^S = D \times D$, according to which no theory is ranked ahead of another, giving the agent absolutely no help in choosing between theories. More generally, (2) 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, (2) does much more than simply limit indifference classes, as we make clear in Section 3.3.

One natural way to ensure that (2) 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 nonsingleton equivalence classes of 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 $\succ^S$ there exists an enumeration $D = \{t_1, t_2, \ldots\}$ such that $t_i \succ^S t_{i+1}$, with strict preference $\succ^S$ occurring for infinitely many

---

[7]In an effort to keep things straight, we use $\succsim$ to denote a relation by which the agent chooses theories and use $\succ$ 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).

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[8]

$$t \succcurlyeq^S t' \quad \Longleftrightarrow \quad C(t) \le C(t') \tag{3}$$

and

$$|C^{-1}(k)| < \infty \quad \forall k \in \mathbb{N}. \tag{4}$$

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

$$\forall h \in H, \quad t \succsim^{LS}_h t' \quad \Longleftrightarrow \quad \begin{cases} \{t \succ^L_h t'\} \\ \text{or} \quad \{t \sim^L_h t' \text{ and } t \succcurlyeq^S t'\}. \end{cases} \tag{5}$$

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 choose only 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 on 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 is falsified. However, the only event that pushes 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 is never pushed out of this class, and eventually is 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 eventually is 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.

---

[8]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 also has a high index according to C_2.

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 that 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 "tries 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 1. Consider a subjective order that ranks any theory that predicts 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 never is 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 (2), noted just after its introduction, is to preclude such cases.   ◇

EXAMPLE 2. Suppose the subjective order is based on a lexicographic accounting of 1's. Theory t is ranked ahead of t' if t predicts a 1 in the first period and t' predicts 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' predicts 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' predicts 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 0's. The theory that corresponds to this outcome ranks below every other possible theory. The agent never reaches this theory and, indeed, predicts 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 (2) is to preclude such possibilities. For every possible theory, there are only finitely many preferred theories under $\succeq^S$, ensuring that exploration guided by $\succeq^S$ eventually hits on the data-generating process (or something observationally equivalent), at which point this theory is 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 they receive 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 on the truth, she soon abandons it in favor of another seemingly equivalent theory. Does it not suffice to assume that the agent sticks to something that has worked in the past?

The phenomenon of inertia, i.e., a preference for the status quo, is familiar from casual observations as well as from psychological studies. Kuhn (1996) argues 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 whether inertia suffices for effective learning, we define the *inertial* relation as that which selects the theory chosen in the previous period if the latter maximizes the likelihood function, and otherwise chooses as does the likelihood relation. Formally, define $\succsim^{LI}$ for all $n > 1$ as

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

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 3. Let D consist of the set of deterministic theories

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

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.

The agent chooses theories from a subset of D. Given history h_n, let $\{t^{h_n, n+k}\}_{k=0}^{\infty}$ be a collection of theories, with the theory $t^{h_n, n+k}$ given by

$$t^{h_n, n+k}(h_n h_l) = \begin{cases} 1 & \text{if } l = k \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Theory $t^{h_n, n+k}$ thus predicts that history h_n will be followed by k periods in which 0 is observed, followed by a 1, and then followed by 0 forever.

Having observed history h_n, the agent's selection rule over theories is given by

$$\mu_{B_{\succsim_{h_n}^{LI}}}(t^{h_n, n+k}) = \frac{1}{2^{k+1}}, \quad k = 0, \ldots. \tag{7}$$

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 period $n + k$. Under this

selection rule, theories that predict a 0 on the next step are equally likely as theories that predict a 1, in accordance with (2).

Suppose the data-generating process is such that $y_n = 0$ for all n, and consider $\succsim^{LI}_{h_n}$ for a history h_n consisting of n 0's. Given (6) and (7), $\succsim^{LI}_{h_n}$ chooses a theory that predicts a sequence of 0's, until it is falsified when the theory predicts its lone 1. The probability that the agent chooses a theory lasting k periods before being falsified is $1/2^{k+1}$. Upon having a theory falsified, the agent must choose a new theory, putting the agent in a position identical to that of the initial period and giving payoff equal to the agent's initial payoff. The agent's payoff is thus given by $\Pi(y_0, \succsim^{LI}) = \frac{1}{2}$.[9] $\Diamond$

The difficulty in this example is that the selection rule over the various sets $B_{\succsim^{LI}_{h_n}}$ invariably 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 $B_{\succsim_h}$.

Under Assumption 3, the agent choosing a theory attaches positive probability to every unfalsified theory in D. This ensures that the agent cannot assuredly avoid selecting the data-generating process, as in Example 3.

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 is not abandoned, and hence the optimal payoff is obtained.

Observe that, in terms of our definitions, the inertial likelihood relation $\succsim^{LI}$ is not purely objective. Specifically, assume that one is given the conditional probabilities of each possible observation given each theory, for every prefix of a history h. This information does not suffice to determine the selection of the relation $\succsim^{LI}$: this relation

---

[9]To calculate this payoff, let v^t_k be the expected average payoff in the next $t \geq 1$ periods, given that the predictor's current theory was correct in the last $k \geq 0$ periods (but not in the $(k+1)$st if there was one). Then

$$v^t_k = \frac{1}{2}\left(\frac{1}{t} \cdot 0 + \frac{t-1}{t}v^{t-1}_0\right) + \frac{1}{2}\left(\frac{1}{t}1 + \frac{t-1}{t}v^{t-1}_{k+1}\right)$$

$$= \frac{1}{t} \cdot \frac{1}{2} + \frac{t-1}{t} \cdot \left(\frac{1}{2}v^{t-1}_0 + \frac{1}{2}v^{t-1}_{k+1}\right),$$

but we also have $v^1_k = \frac{1}{2}$ for all k and hence $v^t_k = \frac{1}{2}$ for all k, t; hence

$$\Pi(y_0, \succsim^{LI}) = \lim_{t \to \infty} v^t_0 = \frac{1}{2}.$$

gives priority to the theory that was chosen in the past, and this past selection is not deterministically implied by the data. Thus, we regard Proposition 3 to be consistent with the message of this paper. Under Assumption 3, $\succsim^{LI}$ leads to effective learning, but the inertial tendency is a subjective consideration.

3.3.4 *Bayesian reasoning*   If the set of conceivable data-generating processes is countable (cf. Assumption 1) and the agent has a strictly positive 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 to be more likely.

There is 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. Either formulation gives a limiting payoff of 1.

4. Random data-generating processes: Likelihood trade-offs

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 idiosyncrasies 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 $\varepsilon \in (0, 1/2)$,

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

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

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

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

$$\log(1 - \varepsilon) = \theta(1) \qquad \text{Maximum possible limiting value.}$$
$$(1 - \varepsilon)\log(1 - \varepsilon) + \varepsilon \log \varepsilon = \theta(1 - \varepsilon) \qquad \text{Value achieved by the data-generating process.}$$
$$\tfrac{1}{2}\log(1 - \varepsilon) + \tfrac{1}{2}\log \varepsilon = \theta(\tfrac{1}{2}) \qquad \text{Value achieved by random choice.}$$

FIGURE 2. Key values of the limiting average log-likelihood function (8).

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,

$$l(t, h_n) = \frac{1}{n}\log(L(t, h_n))$$
$$= \frac{1}{n}\sum_{j=0}^{n-1}\log\big[t(h_j)y_j + (1 - t(h_j))(1 - y_j)\big], \tag{8}$$

which does not converge to zero. We hereafter use "likelihood" to denote the average log likelihood, given by (8).

Let us say that a theory is "correct" in period t if it predicts 1 with probability $1 - \varepsilon$ and 1 occurs, or if it predicts 0 with probability $1 - \varepsilon$ and 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 for proportion p of the time.

A theory that is correct in every period gives 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 achieves payoff $1 - \varepsilon$.[10] Predicting randomly gives likelihood $\theta(\tfrac{1}{2})$ and payoff $\tfrac{1}{2}$. Figure 2 summarizes these observations.

The counterpart of Assumption 1 is now described as follows.

ASSUMPTION 4.

   (i) $D \subset D_\varepsilon$.

  (ii) D is countable.

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

Assumption 4(iii) 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.

---

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

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 2.5 or the subjective relation $\succsim^{LS}$ of Section 3.2. In the random environment, this ensures that the agent eventually *excludes* 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 eventually almost surely has a likelihood value lower than $\theta(1)$, whereas there are always other theories with a likelihood value of $\theta(1)$. That is, insisting on maximum-likelihood theories leads 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 that achieve a likelihood of at least $\theta(\gamma)$ into a single equivalence class.

What would be a good value of γ? We suspect that we should set $\gamma < 1 - \varepsilon$, since any value $\gamma > 1 - \varepsilon$ eventually surely excludes 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 is not ruled out, but there is no guarantee that it is selected. An argument analogous to that establishing Proposition 1 immediately provides the (omitted) proof of the following proposition.

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

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

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

Under the subjective relation, setting $\gamma > 1 - \varepsilon$ again implies that the agent discards the data-generating process as a possible theory and subsequently hops 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,* $\{t \in B_{\succsim_h^{LS,\gamma}} \mid t(h) = 1 - \varepsilon\}$ *and* $\{t \in B_{\succsim_h^{LS,\gamma}} \mid t(h_n) = \varepsilon\}$ *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 that amasses 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 proposition is then immediate.

PROPOSITION 5. *Let Assumptions 2, 4(i) and (ii), 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 is—at least eventually—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 fits 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.

EXAMPLE 4. 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 are ranked at the top of the subjective order: $t_1 \succ^S t_2 \succ^S \cdots \succ^S t_k$ and $t_k \succ^S t'$ for all $t' \notin \{t_1, \ldots, t_k\}$. The essential feature of these theories is that any theory whose likelihood falls short of γ predicts 1 (thereby improving its likelihood). If some of the theories exhibit a likelihood above γ, the one that ranks highest (under $\succ^S$) predicts 0 and the others predict 1.[11]

In each period, at least $k-1$ of the theories $(t_1, \ldots, t_k)$ produces 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 thus is 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

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

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

$$\left[\left(1 - \varepsilon - \tfrac{1}{2}((1 - \varepsilon) - \gamma)\right)(k - 1)\right]n > k\gamma n, \tag{9}$$

or

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

(Since $1 - \varepsilon > \gamma$, such a k exists.) From (9), 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 is used for prediction in every period $n > n^*$ and, by definition, predicts that outcome that appears with probability ε under the data-generating process d. Hence, the agent's payoff converges to ε.   ◇

---

[11]We can formally 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 subhistories 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 $\{t_1, \ldots, t_k\}$ with $l(t_j, h_n) \geq \gamma$. If it is itself, it predicts 0; otherwise, it predicts 1. Observe that each theory $\{t_1, \ldots, t_k\}$ 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 $\{t_1, \ldots, t_k\}$ arises only out of the very last step of the algorithm, in case some of them obtain a likelihood value above the threshold.

It is not important for this example that the theories $(t_1, \ldots, t_k)$ occupy the top k places in the subjective order $\succ^S$. We need only that other theories ranked ahead of any of $(t_1, \ldots, t_k)$ are sufficiently different from d as to have likelihoods that eventually fall (and remain) below those of $(t_1, \ldots, t_k)$.

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 $\gamma \leq 1 - \varepsilon$ and $k \geq 1$ be given. 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

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

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

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

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. In particular, if there are several theories that obtained this likelihood threshold in each of the past $n - k + 1$ periods, the maximizer has to be the 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 achieved at least $\theta(\gamma)$ for at least $(n - k)$ periods out of the past $(n - k + 1)$ periods. If no theory has done as well as $\theta(\gamma)$ for $n - k$ out of the past $n - k + 1$ periods, $\succsim_h^{LS,\gamma k}$ selects the subjectively best ranked among those that achieved at least $\theta(\gamma)$ for at least $(n - k - 1)$ 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 $(\succsim_h^{LS,\gamma k})$. What should be the values of γ and k, and how are they determined?[12] In particular, different values of ε 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 so as to ensure effective learning:

---

[12]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 $1 - \varepsilon$.

PROPOSITION 6. *Under [Assumption 4](#), for every discriminating subjective order $\succ^S$ 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 predict as well as d over any string of k periods becomes negligible. This ensures that the agent eventually uses 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* $j \le n$. Such a theory maximizes the likelihood function for each prefix of the history h_n and, therefore, is 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 simply be endowed with the optimal level of tolerance? Such serendipity is not necessary. 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 $\varepsilon \in (0, 1/2)$ and $d \in D$, almost surely (with respect to the measure over histories induced by d), there exist sequences $\gamma_n \to 1 - \varepsilon$ and $k_n \to \infty$ such that for any $t, t' \in D$,*

$$t \succsim^{S*}_{h_n} t' \iff t \succ^{LS,\gamma_n k_n}_{h_n} t'.$$

The idea behind this result is that the agent can adjust her level of γ in light of her experience. An agent who sets a level of γ too high soon finds 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. As the data accumulate, this boundary level (almost surely) approaches $1 - \varepsilon$. At each step of the process, the relation exploits the 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.

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

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

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

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 discriminating subjective order $\succ^S$, 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 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 simplic-
ity. Simplicity takes the place of our subjective order in these criteria, while our lexico-
graphic criterion is typically replaced by a smooth objective function.

The general form of these measures is

$$\log L(t) - \alpha C(t), \tag{10}$$

where $C(t)$ is a "complexity function" (i.e., a function $C : D \to \mathbb{N}$ that satisfies (4)), and
α is a constant that determines the relative weights placed on the likelihood and on the
complexity of the theory.[13] Gilboa and Schmeidler (2010) offer an axiomatization of this
criterion.

We cannot apply (10), designed to evaluate theories given a fixed set of data, directly
to our setting. As we note, the likelihood $L(t)$ inevitably declines to zero and hence its
log decreases without bound as observations accumulate. This ensures that complex-
ity considerations or any other subjective considerations eventually play no role in the
analysis. We accordingly examine

$$l(t, h) - \alpha C(t), \tag{11}$$

ensuring that likelihood and complexity considerations remain on a common footing.[14]

We can draw a connection between smooth measures such as (11) and our lexico-
graphic criterion. Fix a complexity function $C(t)$ and parameter α, and let $\succsim^{\alpha}$ be the
resulting order over theories induced by (11). How does $\succsim^{\alpha}$ compare to $\succsim^{LS}$, where the
latter is based on the subjective order over theories defined by (5), corresponding to the
complexity function induced by $C(t)$ via (3)?

To simplify the discussion, let us restrict attention to a set of data-generating pro-
cesses $D_{\varepsilon}^{C} \subset D_{\varepsilon}$ with the property that for any $d, d' \in D_{\varepsilon}^{C}$, the average log-likelihood ra-
tio $l(d', h_n)$ converges with probability 1, 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 4, 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*

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

---

[13]For example, the Akaike information criterion (Akaike 1974) 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 1965, 1998, Chaitin 1966, Solomonoff 1964), the minimal
message length criterion (Rissanen 1978, Wallace and Boulton 1968) suggests

$$\log(L(t)) - \mathrm{MDL}(t),$$

where the MDL(t) is the minimum description length of theory t. (See also Wallace (2005) and Wallace and
Dowe (1999).)

[14]In so doing, we move close to criteria such as the Schwarz information criterion (also known as the
Bayesian information criterion (Schwarz 1978)), which retains the additive trade-off, but uses a complexity
measure that depends on the number of observations.

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 do 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(iii)). In addition, we cannot say a priori how small α must be 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. This is just what our lexicographic criterion does. We can accordingly view the lexicographic criterion as the limiting case of the smooth criteria that are 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 and 5.2 for the case of a deterministic data-generating process.

5.1 *Countability*

We have assumed that 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 1979, Chapter 7), and hence the set of conceivable data-generating processes can reasonably be taken to be countable.[15]

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 let T be the set of theories of which the agent can

---

[15]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.

conceive. We may then have a set D that is an (uncountable) superset of T. How does 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 to then 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 $\succ^S$ 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(i) and (iii) (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 $\succ^S$, 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 goes well and the use of the discriminating subjective order $\succ^S$ then ensures a payoff approaching unity whenever the data-generating process is drawn from T. This is also 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.[16] 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

---

[16]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.

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 $T \subset D_0^H$ is unreasonable, as it is tantamount to magically endowing the agent with the ability to solve the celebrated halting problem.[17]

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 $\Pi(d, \succsim) \le 0.5$.*

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 belief period after period. Does 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

$$C(t) \le C(t') \quad \Longleftrightarrow \quad t \succ^S t'.$$

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

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

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

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

and

---

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

(ii) *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 $L(d, h) = L(d', h)$,*

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

Proposition 11 ensures the existence of a computable strategy that yields optimal payoffs, as well as its asymptotic agreement with the (strict part of) the given subjective ordering $\succcurlyeq^S$ over D_0^T.[18] 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 $\succsim$ does not reflect the subjective ranking $\succcurlyeq^S$, but in the long run, any two theories that are equally accurate are 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 regression and nonlinear 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 is 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 *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.[19]

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 are the subject of an immense literature (e.g., Sober 1975). Importantly, ever since Goodman (1954) 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 1965, Chaitin 1966) 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 into each other (see Solomonoff 1964)). We hope that the framework provided in this paper can be used to investigate the role of simplicity in inductive inference.

---

[18]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.

[19]This preference for simplicity has been championed on normative grounds (most famously by William of Occam (see Russell 2004)) and has long been offered as a descriptive model of human reasoning (e.g., Wittgenstein 1922).

Appendix: Proofs

PROOF OF PROPOSITION 1. Assumption 1(iii) 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,

$$B_{\succsim_h^L} = \{d \mid L(d, h) = 1\}.$$

For any finite continuation of h_n, there is a theory $t \in B_{\succsim_h^L}$ 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 predicting $y_n = 1$. 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

$$\Pi(d, \succsim_h^L) = 0.5. \qquad \square$$

PROOF OF PROPOSITION 2. Fix $d \in D$. Since $\succ^S$ is discriminating, there are finitely many theories in $S(d) \equiv \{t \in D \mid t \succ^S d\}$, i.e., that are ranked ahead of or are indifferent to d by the subjective order $\succ^S$. Choose some $t \in S(d)$, and suppose that t and d are not observationally equivalent, meaning that they do not generate identical outcomes $((x_0, y_0), \dots, (x_n, y_n), \dots)$. Then at some period n, theory t is refuted, i.e., the data-generating process produces a history $h_n = ((x_0, y_0), \dots, (x_{n-1}, y_{n-1}), x_n)$ 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 $\succ^S$ 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 $\succsim^L$. $\qquad \square$

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. $\qquad \square$

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 $j = k, \ldots, n$ and (ii) $t \succ^S d$. Hence, for each period $n > k$, there are only a finite number of theories that satisfy conditions (i) and (ii), of which the highest ranked by the subjective order is 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' is reached such that some theory t (possibly d) that satisfies (i) and (ii) is 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

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

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. $\qquad\square$

PROOF OF PROPOSITION 7. The basic idea is to have the agent simulate the choices of theories that correspond to $\succsim^{LS,\gamma k}$ for different values of γ and of k. For values of γ larger than $1 - \varepsilon$, the agent finds 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 finds many 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 approaches $1 - \varepsilon$ and makes 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

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

Given n, define $k = \llcorner n/2 \lrcorner$. For each $\gamma \in \Gamma_n$ and for each $m = k, \ldots, n$, the reasoner finds all the maximizers of $\succsim^{LS,\gamma k}_{h_m}$ (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

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

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 $\succsim^{S*}_{h_n} = D \times D$. In this case, all theories are equivalent in terms of $\succsim^{S*}_{h_n}$, and the reasoner's choice is 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

$$t \succsim^{S*}_{h_n} t' \quad \Longleftrightarrow \quad \begin{cases} t \in M^*(n, \gamma_n) \quad \text{and} \quad t' \notin M^*(n, \gamma_n) \\ \text{or} \quad t, t' \in M^*(n, \gamma_n) \quad \text{and} \quad t \succcurlyeq^S t' \\ \text{or} \quad t, t' \notin M^*(n, \gamma_n). \end{cases}$$

That is, the $\succcurlyeq^S$-most-preferred 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 $\gamma \nearrow 1 - \varepsilon$. $\qquad\square$

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 $\theta(1) - \alpha C(d) < \theta(\varepsilon) - \alpha C(\hat{d})$ is ever be chosen by the relation $\succsim^\alpha$, no matter what the history.

The agent's choice of theory in each period thus is drawn from the finite set $D_\varepsilon^C(\alpha) \equiv \{d \in D_\varepsilon^C \mid \theta(1) - \alpha C(d) < \theta(\varepsilon) - \alpha C(\hat{d})\}$.

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

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

But since $d^* \in D_\varepsilon^C(\alpha)$ 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. $\square$

PROOF OF PROPOSITION 9. The relation $T \times T$ guarantees a random choice (by Assumption 1(iii)), 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, $\succsim^{LS,\varepsilon}$ eventually selects d or a theory equivalent to d. Predictions subsequently are perfect, ensuring that $\succsim^{LS,\varepsilon}$ does 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 $d \in T$. $\square$

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 results in a payoff of 0. The maximal payoff for the agent at each period is 0.5, obtained by the random prediction $t_h(h) = 0.5$. $\square$

PROOF OF PROPOSITION 11. The basic idea is to construct the relation $\succsim$ by combining the underlying subjective order $\succ^S$ with the time complexity of the machine.

Let $D_0^T = \{t_1, t_2, \ldots\}$ 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

$$C(t) \le C(t') \quad \Longleftrightarrow \quad t \succcurlyeq^S t'.$$

Define $c : D \times H \to \mathbb{N} \cup \{\infty\}$ to be the length of computation, that is, $c(t, h) \in \{1, 2, \ldots, \infty\}$ 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 jth prefix of h. Using this function, we define our candidate relation over theories:

$$t' \succsim_h t \quad \Longleftrightarrow \quad \begin{cases} L(t', h) > L(t, h) \\ \text{or} \quad [L(t', h) = L(t, h) \text{ and } C^*(t', h) \le C^*(t, h)]. \end{cases}$$

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 has likelihood 1. 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 $\succcurlyeq^S$, we have $C(t) < \infty$ and hence $C^*(t, h) < \infty$.

Given $C^*(t, h)$, there are finitely many machines t' with $C(t') \le C^*(t, h)$, 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 $d \in D_0^B$, $c(t, h_n) \le K(d)$ and

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

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(ii).

We now turn to 11(i), 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

Akaike, Hirotugu (1974), "A new look at the statistical model identification." *IEEE Transactions on Automatic Control*, 19, 716–723. [204]

Al-Najjar, Nabil I. (2009), "Decision makers as statisticians: Diversity, ambiguity, and learning." *Econometrica*, 77, 1371–1401. [186]

Chaitin, Gregory J. (1966), "On the length of programs for computing binary sequences." *Journal of the Association for Computing Machinery*, 13, 547–569. [204, 208]

Gilboa, Itzhak and David Schmeidler (2010), "Simplicity and likelihood: An axiomatic approach." *Journal of Economic Theory*, 145, 1757–1775. [204]

Goodman, Nelson (1954), *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, Massachusetts. [208]

Hopcraft, John E. and Jeffrey D. Ullman (1979), *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, Massachusetts. [205]

Kolmogorov, Andrei N. (1965), "Three approaches to the quantitative definition of information." *Problems of Information Transmission*, 1, 1–7. [204, 208]

Kolmogorov, Andrei N. (1998), "On tables of random numbers." *Theoretical Computer Science*, 207, 387–395. [204]

Kuhn, Thomas S. (1996), *The Structure of Scientific Revolutions*, third edition. University of Chicago Press, Chicago. [194]

Rissanen, Jorma (1978), "Modelling by shortest data description." *Automatica*, 14, 465–471. [204]

Russell, Bertrand (2004), *History of Western Philosophy*. Routledge, London. [208]

Schwarz, Gideon (1978), "Estimating the dimension of a model." *Annals of Statistics*, 6, 461–464. [204]

Sober, Elliott (1975), *Simplicity*. Clarendon Press, Oxford. [208]

Solomonoff, Ray J. (1964), "A formal theory of inductive inference I, II." *Information Control*, 7, 1–22, 224–254. [204, 208]

Vapnik, Vladimir N. (1998), *Statistical Learning Theory*. Wiley, New York. [186]

Vapnik, Vladimir N. and Alexey Y. Chervonenkis (1971), "On the uniform convergence of relative frequencies of events to their probabilities." *Theory of Probability and Its Applications*, 16, 264–280. [186]

Wallace, Christopher S. (2005), *Statistical and Inductive Inference by Minimum Message Length*. Springer, New York. [204]

Wallace, Christopher S. and D. M. Boulton (1968), "An information measure for classification." *The Computer Journal*, 11, 185–194. [204]

Wallace, Christopher S. and David L. Dowe (1999), "Minimal message length and Kolmogorov complexity." *The Computer Journal*, 42, 270–283. [204]

Wittgenstein, Ludwig (1922), *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, London. [208]