Second-order induction in prediction problems

Rossella Argenziano[a,1] **and Itzhak Gilboa**[b,c,1,2]

[a]Department of Economics, University of Essex, Colchester CO4 3SQ, United Kingdom; [b]Economics and Decision Sciences Department, École des Hautes Études Commerciales de Paris, 78351 Jouy-en-Josas Cedex, France; and [c]Berglas School of Economics, Tel-Aviv University, Tel Aviv 6997801, Israel

Edited by Matthew O. Jackson, Stanford University, Stanford, CA, and approved April 9, 2019 (received for review January 28, 2019)

Agents make predictions based on similar past cases, while also learning the relative importance of various attributes in judging similarity. We ask whether the resulting "empirically optimal similarity function" (EOSF) is unique and how easy it is to find it. We show that with many observations and few relevant variables, uniqueness holds. By contrast, when there are many variables relative to observations, nonuniqueness is the rule, and finding the EOSF is computationally hard. The results are interpreted as providing conditions under which rational agents who have access to the same observations are likely to converge on the same predictions and conditions under which they may entertain different probabilistic beliefs.

belief formation | empirically optimal similarity function | learning | kernel estimation | generalized context model

Where do beliefs come from? How do economic agents predict future realizations of relevant variables? We consider an agent who, in each period, predicts the realization of a variable of interest, after observing the realization of presumably related other variables. Agents predict that the variable of interest will be a weighted average of its past values, and they assign a higher weight to values that were observed under more similar circumstances. This method is known in statistics and machine learning as "kernel estimation" (see refs. 1 and 2 as well as support vector machines in refs. 3 and 4). Surprisingly, a very similar formula appeared in the psychological literature in the context of the generalized context model (GCM) (5, 6). The latter deals with a classification task, where participants are asked to decide to which category an object belongs. The GCM suggests that the category chosen is the "most frequent" one encountered, where frequency is weighted by similarity.

While psychology aims at modeling human reasoning, whether optimal or not, statistics and machine learning attempt to develop effective ways of prediction based on past data, with no claim to describe the way people think. A priori, there is no reason to believe that these disciplines would converge to the same class of models. The fact that they did independently derive similar techniques makes these techniques very promising for modeling beliefs of economic agents. As noted by ref. 7 (p. 831), "...kernel methods have neural and psychological plausibility, and theoretical results concerning their behavior are therefore potentially relevant for human category learning." This paper presents a model of belief formation based both on insights from the GCM and on kernel techniques.

The GCM assumes that individuals store "exemplars" (objects) in their memory as points in a multidimensional psychological space, in which each dimension is a feature of the objects. They then classify new objects based on their similarity to the stored exemplars (see ref. 8 for a survey). Individuals use selective-attention weights to measure the importance of each feature in their similarity assessments. The empirical evidence reviewed by ref. 9 strongly suggests that the similarity between two objects is measured as a negative exponential function of their distance in this psychological space. Crucially for our model, experimental evidence shows that individuals use different selective-attention weights for different tasks and, moreover, that for any given task they learn the weights that optimize their classification performance in that context (6, 10, 11).

Inspired by these results on classification tasks, we present a model of prediction based on two levels of learning. (We refer here to the learning needed to form prior beliefs and not to Bayesian learning that such beliefs may later be used for.) First, we assume that the value of a variable y is estimated by the similarity-weighted average of its past realizations. Specifically, observation i consists of a realization of a vector of predictors x_i and a value of the predicted variable y_i; a new datapoint x_p is presented, and the task is to estimate the value of the corresponding y_p. First-order induction assumes a similarity function $s(x_i, x_p) \geq 0$ such that y_p is estimated by the $s(x_i, x_p)$ -weighted average of past y_is. Past occurrences are weighted by their similarity: Values y_i observed under circumstances x_i more similar to the current x_p gain higher weight. In statistical terms, $\overline{y}_p^s$ is the kernel-based estimate of y_p with kernel s. Following the empirical regularity observed by ref. 9, we use a similarity function that is a negative exponential of the weighted distance between pairs of vectors of predictors. The weights given to the different predictors are analogous to the selective-attention weights of the GCM in that they identify the relative importance of each component in the similarity assessment.

The second level of learning involves finding the optimal weights. We model this problem by a leave-one-out cross-validation technique and refer to a similarity function that uses optimal weights as an empirically optimal similarity function (EOSF). [Ref. 12 also suggested the notion of "empirical similarity," based on the notion of a maximum-likelihood estimator of the similarity, assuming that the actual data-generating process (DGP) is similarity based. Refs. 13–16 analyzed the asymptotic properties of such estimators. The asymptotic results in this literature assume a given DGP (typically, using a formula such as **[1]**, with a noise variable, as the "true" statistical model), whereas our

Significance

How do people generate beliefs about economic, political, and social events? A natural formula for the predicted value of a variable is its weighted average value in the past, where past values are given a higher weight if they were observed in circumstances more similar to the current ones. Agents learn from the data the best way to assess the similarity of past cases to the present one. The basic formula and this learning process appeared both in statistics and in psychology and they thus make sense for modeling economic agents. We study this process and identify circumstances under which agents are likely to agree on predictions and conditions under which disagreement over predictions may be reasonably expected.

Author contributions: R.A. and I.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]R.A. and I.G. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: tzachigilboa@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1901597116/-/DCSupplemental.

Published online May 7, 2019.

results are more agnostic about the underlying DGP.] Because this process deals with learning how first-order induction should be performed, it is dubbed second-order induction. In statistical terms, this is akin to finding the optimal kernel to estimate y_p (17).

We conceive of this two-stage learning process as an idealized model of the way economic agents form beliefs and we ask whether rational individuals with access to the same information will agree on their predictions. We investigate whether the EOSF is unique and easily computable. If that is the case, we can expect agents to agree on the similarity function to be used and consequently to share the same predictions. We find that, if the number of predictors is fixed, and the predicted variable is a function of the predictors, then, as the number of observations grows following an i.i.d. process, the EOSF will learn the functional relationship. The EOSF will be almost unique with high probability, with different such functions providing similar predictions (*Proposition 1*). By contrast, if the number of predictors is large relative to the number of observations, it is highly probable that the EOSF will not be unique (*Proposition 2*). Further, if the number of predictors is not bounded, the problem of finding the EOSF is nondeterministic polynomial-time complete (NPC) (*Theorem 1*).

Our results suggest that whether rational agents who have access to the same information will agree on their predictions depends, to a large extent, on the comparison of the number of potentially relevant variables and the number of observations. Consider two prediction problems: In the first one, an agent tries to estimate the probability of water boiling. In the second one, an agent tries to estimate the probability of success of a revolution attempt. In the first problem, the number of observations can be increased at will, through experimenting, and there are a relatively limited number of variables to take into account, such as temperature, pressure, and a few other experimental conditions. In this type of problem it stands to reason that the EOSF be unique. Further, as the number of variables is not large, the computational complexity result has little bite. Thus, different people are likely to come up with the same similarity function and therefore with the same probabilistic predictions. By contrast, in the revolution example the number of observations is very limited. One cannot gather more data at will, either by experimentation or by empirical research. To complicate things further, the number of variables that might be relevant predictors is very large: Researchers may come up with novel perspectives on a given history and suggest new potentially relevant military, economic, and sociological variables. In this type of example our results suggest that the EOSF may not be unique and that, even if it is unique, people may fail to find it. As a result, it may not be too surprising that experts may disagree on the best explanation of historical events and, consequently, on predictions for the future.

Model

Case-Based Beliefs. The basic problem we deal with is predicting a value of a variable $y \in \mathbb{R}$ based on other variables $x^1, \ldots, x^m \in \mathbb{R}$. We assume that there are n observations of the values of the x variables and the corresponding y values and, given a new value for the xs, attempt to predict the value of y. We use the terms "cases" and "similarity," as equivalent to "observations" and "kernel."

We assume that prediction is made based on a similarity function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$. Such a function is applied to the observable characteristics of the problem at hand, $x_p = (x_p^1, \ldots, x_p^m)$, and the corresponding ones for each past observation, $x_i = (x_i^1, \ldots, x_i^m)$, so that $s(x_i, x_p)$ would measure the degree to which the past case is similar to the present one. The similarity function should incorporate not only intrinsic similarity judgments, but also judgments of relevance, recency, and so forth.

More formally, let the set of predictors be indexed by $j \in M \equiv \{1, \ldots, m\}$ for $m \geq 0$. When no confusion is likely to arise, we refer to the predictor as a "variable" and also refer to the index as designating the variable. The predictors $x \equiv (x^1, \ldots, x^m)$ assume values (jointly) in $\mathbb{R}^m$ and the predicted variable, y–in $\mathbb{R}$. The prediction problem is defined by a pair (B, x_p) where $B = ((x_i, y_i))_{i \leq n}$ (with $n \geq 0$) is a database of past observations (or cases), $x_i = (x_i^1, \ldots, x_i^m) \in \mathbb{R}^m$, and $y_i \in \mathbb{R}$, while $x_p \in \mathbb{R}^m$ is a new data point. The goal is to predict the value of $y_p \in \mathbb{R}$ corresponding to x_p.

Given a function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$, the value of y_p is estimated by the similarity-weighted average formula

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)}. \quad [1]$$

In the case $s(x_i, x_p) = 0$ for all $i \leq n$, we set $\overline{y}_p^s = y_0$ for an arbitrary value $y_0 \in \mathbb{R}$. (We choose some value y_0 only to make the expression $\overline{y}_p^s$ well defined. Its choice will have no effect on our analysis.) This formula is identical to the kernel-averaging method (where the similarity s plays the role of the kernel function). Similarity-weighted estimation as in **[1]** was axiomatized in refs. 18 and 19.

We use the similarity function

$$s^w(x, x') = \exp\left(-\sum_{j=1}^m w^j \left(x^j - x'^j\right)^2\right) \quad [2]$$

with $w_j \geq 0$. [Our results hold for other similarity functions as well. First, the distance it is based on may be based on any seminorm n_w, such as $n_w(x, x') = \left(\sum_{j=1}^m w^j \left|x^j - x'^j\right|^r\right)^{1/r}$ for $r \geq 1$. (Note that these are seminorms because some w^js may vanish.) The key feature we need is that $n_w(x, x') = 0$ iff $x^j = x'^j$ for all j such that $w^j > 0$. Second, one can select other decreasing functions (rather than the exponential), as long as they vanish at ∞.]

Similarity functions that are negative exponentials of norms on the Euclidean space were axiomatized by ref. 20. Refs. 12 and 19 specified the norm to be a weighted Euclidean distance. We use the extended nonnegative reals, $\mathbb{R}_+ \cup \{\infty\} = [0, \infty]$, allowing for the value $w^j = \infty$. Setting w^j to ∞ would be understood to imply $s^w(x, x') = 0$ whenever $x^j \neq x'^j$, but if $x^j = x'^j$, the jth summand in **[2]** will be taken to be zero. In other words, we allow for the value $w^j = \infty$ with the convention that $\infty \cdot 0 = 0$. For the computational model, the value ∞ will be considered an extended rational number, denoted by a special character (say "∞"). The computation of $s^w(x, x')$ first goes through all $j \leq m$, checking if there is one for which $x^j \neq x'^j$ and $w^j = \infty$. If this is the case, we set $s^w(x, x') = 0$. Otherwise, the computation proceeds with **[2]** where the summation is taken over all js such that $w^j < \infty$.

Empirically Optimal Similarity Function. How do individuals select a similarity function? Evidence in refs. 6, 10, and 11 supports the notion that individuals learn the weights that optimize their performance in a classification task. The notion of second-order induction is designed to capture this idea in the context of estimation. It suggests that, within a given class of possible functions, $\mathcal{S}$, individuals choose one that fits the data best. [Note that the axiomatic derivations mentioned above (18–20) rely on the implicit assumption that the similarity function does not change from one prediction problem to the next. It is natural, however, to think of first- and second-order induction occurring at different time scales. The assessment of y based on x values occurs continuously, while learning of the similarity function

occurs relatively infrequently. Thus, the axiomatic derivations hold approximately, and the appropriate similarly function is learned over longer time spans.

To what extent does a function "fit the data"? One popular technique to evaluate the degree to which a prediction technique fits the data is the "leave-one-out" cross-validation technique: For each observation i, one may ask what would have been the prediction for that observation, given all of the other observations, and use a loss function to assess the fit. In our case, for a database $B=((x_i,y_i))_{i\leq n}$ and a similarity function s, we simulate the estimation of y_i, if only the other observations $((x_k,y_k))_{k\neq i}$ were given, using the function s; the resulting estimate is compared with the actual value of y_i, and the similarity is evaluated by the mean-squared error it would have had.

Explicitly, we consider the set of similarity functions $S=\{\,s^w \mid w\in[0,\infty]^m\,\}$. For $w\in[0,\infty]^m$, let

$$\overline{y}_i^s=\frac{\sum_{k\neq i}s^w(x_k,x_i)y_k}{\sum_{k\neq i}s^w(x_k,x_i)}$$

if $\sum_{k\neq i}s^w(x_k,x_i)>0$ and $\overline{y}_i^s=y_0$ otherwise. Define the mean-squared error to be

$$MSE\,(w)=\frac{\sum_{i=1}^n(\overline{y}_i^s-y_i)^2}{n}.$$

[Analogous results hold for other loss functions (such as the average absolute value of the deviations) and other cross-validation techniques, as long as they yield 0 loss if, and only if, a perfect fit is obtained in-sample.] We also assume that there is a preference for using fewer variables rather than more. A variable with weight $w^j>0$ incurs some fixed cost associated with managing it, collecting the data, recalling it, etc. Thus, in a way that parallels the "adjusted R^2" in regression analysis, we define the adjusted MSE by $AMSE(w,c)\equiv MSE(w)+c|supp\,(w)|$, where $supp\,(w)\equiv\{\,j\leq m \mid w^j>0\,\}$ and $c>0$. We also use $supp(A)$ to denote the set of supports of all of the weight vectors in A.

We intuitively think of an EOSF as a function s^w that minimizes the $AMSE$, but we need to be careful: The argmin of the $AMSE$ may be empty:

Observation 1. *There are databases and $c_0>0$ such that, for every $0<c<c_0$,*

$$\arg\min_{w\in[0,\infty]^m}AMSE\,(w,c)=\arg\min_{w\in[0,\infty]^m}MSE\,(w)=\varnothing.$$

(*Observation 1* is proved in *SI Appendix*). The reason that the argmin of the MSE, and hence of the $AMSE$, may be empty is that the MSE is well defined at $w^j=\infty$ but need not be continuous there. We are therefore interested in vectors w that obtain the lowest $AMSE$ approximately. More precisely, we define the ε-empirically optimal similarity function as follows:

Definition 1. *For $\varepsilon>0$, a function s^w is an ε-empirically optimal similarity function (ε-EOSF) if*

$$w\in\varepsilon\text{-}\arg\min AMSE=$$
$$\left\{w\in[0,\infty]^m \mid AMSE\,(w,c)\leq\inf_{w'}AMSE\,(w',c)+\varepsilon\right\}.$$

The ε-$\arg\min AMSE$ is, thus, the set of weight vectors that are ε-optimal. We are interested in the shape of this set for small $\varepsilon>0$. We informally use the terms "an EOSF" to refer to a 0-EOSF, if such exists, and to an ε-EOSF for a small ε if not, as will be clear from the context.

Results

Almost Uniqueness. In this section we provide three results. Their proofs are contained in *SI Appendix*. We first consider the case in which there is an underlying functional relationship between y and x, such that for some function f we have $y=f\,(x)$. This implies, in particular, that y_i depends only on (x_i) and not on past values of x or of y itself. The agents do not need to know or assume that such a relationship exists, but we would expect that, with sufficiently many observations that represent the entire domain, they would figure it out. This is indeed the message of the following result.

Assume that the observations (x_i,y_i) are i.i.d. For simplicity, assume also that each x_i^j and each y_i are in the bounded interval $[-K,K]$ for $K>0$. Let g be the joint density of x, with $g\,(x)\geq\eta>0$ for all $x\in X\equiv[-K,K]^m$, and let a continuous $f:X\to[-K,K]$ be the underlying functional relationship between x and y so that $y_i=f\,(x_i)$. [A similar result would hold if we allow y_i to be distributed around $f\,(x_i)$ with an i.i.d. error term.] Refer to this data-generating process as (g,f). Given such a process, we say that a variable x^j is informative if f is not constant with respect to x^j and denote by $I(f)$ the set of indexes j such that x^j is informative.

Proposition 1. *Assume a data-generating process (g,f) (where f is continuous). Let there be given $\nu,\xi>0$. There are an integer N and $W\geq0$ such that for every $n\geq N$, for any vector w such that $W\leq w^j<\infty$ for all $j\leq m$ we have*

$$P\,(MSE\,(w)<\nu)\geq1-\xi,$$

where the probability $P=P\,(n,m,g,f)$ is the measure induced by the process described above.

Conversely, if $j\in I(f)$, then for every $W\geq0$ and $\xi>0$ there exist $\nu>0$ and N such that, if $w^j\leq W$, then, for every $n\geq N$,

$$P\,(MSE\,(w)>\nu)\geq1-\xi.$$

Consequently, for every $\xi>0$ there exist N and $c_0>0$ such that, for every $n\geq N$, and every $c<c_0,0<\varepsilon<c/2$,

$$P\,(w\in\varepsilon\text{-}\arg\min AMSE\implies supp(w)=I(f))\geq1-\xi.$$

Proposition 1 deals with the case that y_i is a continuous function of x_i, fixed for all observations. Thus, the question is whether an agent who thinks in terms of similar cases will be able to predict y given x without knowing or even conceiving of such a function.

Proposition 1 addresses this question by two statements and a corollary. On the positive side, it guarantees that if the weights attached to all variables are high enough (but finite) and there are sufficiently many observations, then, with very high probability, the MSE will be small. This is consistent with known results about convergence of kernel estimation techniques (1, 2, 17) although we are unaware of a statement of a result that directly implies this one. On the other hand, the second part of *Proposition 1* states that, if the weight on an informative variable x^j is bounded, then the MSE will be bounded from below. Finally, as a result, with very high probability, all of the weight vectors in ε-$\arg\min AMSE$ share the same support, namely the set of informative variables.

Denoting a "ball" of ∞ as $N_W\,(\infty)=\{\,w\in[\,0,\infty)^m \mid w^j\geq W\;\forall j\leq m\}$, the first part of *Proposition 1* states that, given (a small) $\nu>0$, there exists (a large) W such that any point in $N_W\,(\infty)$ is, with high probability, in ν-$\arg\min MSE$; the second part states that, given (a large) W, there exists (a small) $\nu>0$ such that any point in ν-$\arg\min MSE$ is, with high probability, in $N_W\,(\infty)$.

These first two parts of *Proposition 1* jointly establish that the ε-EOSF is "almost unique." Clearly, uniqueness in its literal sense cannot be expected, as we do not consider the $\arg\min AMSE$ (which may be empty) but the ε-$\arg\min AMSE$. However, *Proposition 1* states that this optimal set is closely related to neighborhoods of infinity, $N_W(\infty)$. In bold strokes, the informative variables would be identified by the ε-EOSF as having a high weight w^j. Hence, under the conditions of *Proposition 1* different individuals who use nearly optimal similarity functions are likely to converge to similar beliefs. If $y=f(x)$ and if we assume, for simplicity, that f depends on all variables, then such individuals may assign different weights to the variables in their similarity functions, but they should all be rather large weights. As a result, in predicting any given y_p they would use only past observations with x_i values that are very close to x_p for making predictions. Given continuity of $f(x)$, their predictions will not vary significantly.

The paradigmatic example in which *Proposition 1* applies is experimentation. If reality is simple enough to have $y=f(x)$, and one can conduct many independent experiments for a variety of x values, one would learn the relationship without needing to assume that a functional relationship exists or to state the findings in the language of such a relationship. Using an ε-EOSF would be enough to guarantee that the agent makes predictions as if she realized that the functional relationship existed. *Proposition 1* can thus explain how different agents converge on the belief that water boils at 100 °C, with some corrections for the air pressure, but disregarding other variables such as the identity of the person who conducts the experiment. Assume instead that the agents are interested in the possibility of a revolution or a financial crisis. The number of relevant observations is rather limited. One cannot run experiments on revolutions. Moreover, the phenomenon of interest is highly complex, and a large variety of variables might a priori be relevant to its prediction. Thus, rather than thinking of n as large relative to m, we consider the opposite case, in which there are many variables relative to observations.

Formally, given n, m, assume that for each $i \leq n$, y_i is drawn, given $(y_k)_{k<i}$, from a continuous distribution on $[-K, K]$ with a continuous density function h_i bounded below by $\eta > 0$. Let v be a lower bound on the conditional variance of y_i (given its predecessors). Next assume that, for every $j \leq m$ and $i \leq n$, given $(y_i)_{i\leq n}$, $\left(x_i^l\right)_{i\leq n, l<j}$, and $\left(x_k^j\right)_{k<i}$, x_i^j is drawn from a continuous distribution on $[-K, K]$ with a continuous conditional density function g_i^j bounded below by $\eta > 0$. Thus, we allow for a rather general class of data-generating processes, where, in particular, the xs are not constrained to be independent. (The assumption of independence of the y_is is used only to guarantee that each observation y_i has sufficiently close other observations, and it can therefore be significantly relaxed.)

The message of the following result is that as the number of observations, n, grows, if the number of variables, m, grows sufficiently fast, then the ε-EOSF is nonunique in a fundamental way: There are weight vectors in the ε-$\arg\min AMSE$ that assign positive weight to distinct sets of variables, but not to their union. The fact that the ε-$\arg\min AMSE$ is not a singleton is hardly surprising, as we allow the $AMSE$ to be ε-away from its infimum and thus expect the ε-$\arg\min AMSE$ to be a set of weights w with a nonempty interior. Indeed, this was found to be the case even under the conditions of *Proposition 1*, which we interpret as a learning result of an almost-unique similarity function. But the following proposition suggests that, assuming a process as discussed here, the nonuniqueness of the weights of the ε-EOSF is not a matter of approximations. More precisely, the set of all supports of the weight vectors in the ε-$\arg\min AMSE$ will typically not be closed under union. For example, we might find one ε-EOSF whose weight vector has a support $J \subset M$ and another such function whose corresponding support is a distinct J', while no ε-EOSF assigns positive weights to all of the variables in $J \cup J'$. Hence, agents who seek an ε-EOSF to explain the data may believe either that J is the set of predictors or that J' is, but none would adopt both sets.

Proposition 2. *Let there be given $c \in (0, v/2)$. There exists $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$ and for every $\delta > 0$ there exists $N = N(c, \varepsilon, \delta)$ such that for every $n \geq N$ there exists $M(n)$ such that for every $m \geq M(n)$, denoting by $P = P\left(n, m, (h_i)_{i\leq n}, \left(g_i^j\right)_{j\leq m, i\leq n}\right)$ the measure induced by the process described above,*

$$P\left(supp(\varepsilon\text{-}\arg\min AMSE) \text{ is not closed under union}\right) \geq 1 - \delta.$$

Proposition 2 suggests a result that is, in a sense, the opposite of *Proposition 1*: The latter proved that, with very high probability, the ε-EOSF will be almost unique, with the support of the EOSF weight vectors including all informative variables; the present result shows that, with very high probability, the supports of the weight vectors of the ε-EOSFs will include distinct sets of variables but not their union. (The proof shows that these sets can also be disjoint.)

Which assumptions are responsible for these starkly different conclusions?

Two main differences arise when comparing the conditions of the two propositions: First, *Proposition 1* assumes that there exists an underlying functional relationship f between x and y, such that each y_i depends only on the observed x_i. Thus, there is something to be learned. And, indeed, the reason that different ε-EOSFs need to be close to each other, or at least to provide close predictions, is that they all uncover the same "truth." By contrast, no such underlying relationship is assumed in *Proposition 2*. Thus, convergence to the truth cannot serve as an engine of agreement.

Second, the order of quantifiers is reversed in the two propositions: In *Proposition 1* it is assumed that the number of predictors, m, is fixed, and the number of observations is driven to infinity. By contrast, *Proposition 2* assumes almost the opposite. True, the number of observations, n, is not held fixed; but the number of variables grows relative to n. (Holding n fixed, a perfect fit for the y_is will not be obtained even if m grows to infinity.) Thus, uniqueness (as in *Proposition 1*) is possible because there are relatively many observations and few variables, and it is impossible (in *Proposition 2*) if the converse is true.

Intuitively, the reason that *Proposition 2* holds is that, with a large set of randomly drawn variables, there is a high probability that a subset thereof (and even a single one) would provide a near-perfect fit. As this holds for any large enough set of variables, there will be disjoint sets that provide near-perfect fit, and thus the ε-EOSF will be nonunique in a way that we think of as "fundamental."

Complexity. *Proposition 2* suggests one reason why rational agents faced with the same prediction problem might adopt similarity functions with very different weights and therefore disagree in their predictions. In this subsection, we present a second reason why this may occur: As the number of possible predictors in a database grows, so does the complexity of finding the ε-EOSF, even if it is almost unique. Formally, we define the following problem.

Problem 1 (ε-EOSF). *Given integers $m, n \geq 1$, a database of rational valued observations, $B = ((x_i, y_i))_{i\leq n}$, and (rational) numbers $c, R \geq 0$, is there a vector of extended rational nonnegative numbers w such that $AMSE(w, c) \leq R$?*

And we can state the following:

Theorem 1. *ε-EOSF is NPC.*

Theorem 1 states that *Problem 1* is computationally hard: There is no known algorithm that can solve it in polynomial time.

It follows that, when many possibly relevant variables exist, as in the case of predicting a social phenomenon, we should not assume that people can find an (or the) ε-EOSF.

The key assumption that drives the combinatorial complexity is that there is a fixed cost associated with including an additional variable in the similarity function. That is, the $AMSE$ is discontinuous at $w^j = 0$. This discontinuity at 0 adds the combinatorial aspect to the $AMSE$ minimization problem and allows the reduction of combinatorial problems used in our proof. *Theorem 1* does not directly generalize to an objective function that is continuous at zero and it is possible that it does not hold in this case. [See also ref. 21, which finds that the fixed cost for including a variable is the main driving force behind the complexity of finding an optimal set of predictors in a regression problem (as in ref. 22).]

Second-Order Induction and Learnability. Our analysis can be viewed as adding to a large literature on what can and what cannot be learned. We consider the problem of predicting y_p based on a database $(x_i, y_i)_{i \leq n}$ and the value of x_p allowing for three types of setups:

i) There exists a basic functional relationship, $y = f(x)$, where one may obtain observations of y for any x one chooses to experiment with.
ii) There exists a basic functional relationship, $y = f(x)$, and one may obtain i.i.d. observations (x, y), but cannot control the observed xs.
iii) There is no bounded set of variables x such that y_i depends only on x_i, independently of past values.

Setup *i* is the gold standard of scientific studies. It allows testing hypotheses, distinguishing among competing theories, and so forth. However, many problems in fields such as education or medicine are closer to setup *ii*. In these problems one cannot always run controlled experiments, be it due to the cost of the experiments, their duration, or the ethical problems involved. Still, statistical learning is often possible. The theory of statistical learning (23) suggests the Vapnik–Chervonenkis (VC) dimension of the set of possible functional relationships as a litmus test for the classes of functions that can be learned and those that cannot. Finally, there are problems that are closer to setup *iii*. The rise and fall of economic empires, the ebb and flow of religious sentiments, social norms, and ideologies are all phenomena that affect economic predictions, yet do not belong to problems of type *i* or *ii*. In particular, there are many situations in which there is causal interaction among different observations, as in autoregression models. In this case we cannot assume an underlying relationship $y = f(x)$, unless we allow the set of variables x to include past values of y, thereby letting m grow with n.

Our positive learning result (*Proposition 1*) assumes that there is an underlying functional relationship of the type $y = f(x)$, keeps m fixed, and lets n grow to infinity, as in setup *ii*. However, it does not assume that the predictor is aware of the existence of such a function, nor that she tries to learn it by selecting the best fit from a given class $\mathcal{F}$ of functions of x. Rather, she predicts y by averaging over its past values, as in kernel regression (1, 2). Indeed, *Proposition 1* is in the spirit of ref. 17 in showing that, as n grows, kernel estimation with optimal kernel parameters leads to good predictions. However, ref. 17 and the bulk of the literature that followed focus on a single parameter, the kernel's bandwidth. In our model, a separate parameter is learned for each variable: Agents learn which variables to attend to. In this context, *Proposition 1* might be viewed as saying that this additional freedom does not come at the expense of the optimality in the results of ref. 17.

Our negative result (*Proposition 2*) may sound familiar: With few observations and many variables, learning is not to be expected. However, our notion of a negative result is starker than that used in the bulk of the literature: We are not dealing with failures of convergence with positive probability, but with convergence to multiple limits. In particular, we conclude that, with very high probability, there will be vastly different similarity functions, each of which obtains a perfect fit to the data. When applied to the generation of beliefs by economic agents, our result discusses the inevitability of large differences in opinion.

Finally, our complexity result (*Theorem 1*) points at a different difficulty: The task of finding the ε-EOSF is computationally complex. There is no known algorithm that can find it in polynomial time. Thus, even if the process is learnable in the sense of being governed by a function from a low VC-dimension class, agents using first- and second-order induction for their predictions might still not be able to learn it correctly.

Discussion

Comparison with Regression. Similar results hold for linear regression. It is well known that if the underlying DGP is such that y is a linear function of x (with random noise), the ordinary least-squares (OLS) method would uncover the relationship when n is large; that if, by contrast, m is larger than n, then, generically, there will be multiple sets of variables that obtain a perfect fit to the data; and also that finding the best set of predictors is nondeterministic polynomial-time hard (22).

There are, however, important differences between the models. First, overfitting is not a problem for the similarity model discussed here as it is for regression analysis. For example, for a fixed number of observations, n, the number of predictors, m, can go to infinity without obtaining a perfect fit. The reason is that, as opposed to regression analysis, in our model y cannot be predicted as a function of the x variables directly. It is predicted only as a function of other y values, where the x values mediate this relationship via the similarity weights. To consider a stark example, if the database consists of only two observations, with $y_1 = 0$ and $y_2 = 1$, we obtain $MSE = 1$ for any set of predictors, irrespective of how large m is and of the values of these xs. (This is also the reason that *Proposition 2* required a large n before demanding that m be large relative to n.)

Second, OLS learning works well if the underlying relationship is indeed linear. More generally, many learning methods work well if the DGP belongs to a particular domain. By contrast, our learning process assumes very little about the true DGP, thus allowing agents to learn a variety of processes. One could argue that, on top of its simplicity, this is a significant advantage from an evolutionary viewpoint.

Compatibility with Bayesianism. There are several ways in which the learning process we study can relate to the Bayesian approach. First, one may consider our model as describing the generation of prior beliefs, along the lines of the "small world" interpretation of the state space (as in ref. 24, section 5.5).

Alternatively, one can adopt a "large world" or "grand state space" approach, in which a state of the world resolves all uncertainty from the beginning of time, and a prior is defined over the space of all such states. This approach is also compatible with the process we describe, when the prior beliefs assign high probability to the data-generating process being governed by a similarity function. In the context of equilibrium selection in a coordination game (such as a revolution), second-order induction may thus define a natural focal point that Bayesian players would find optimal to adhere to.

Agreement. Economic theory tends to assume that, given the same information, rational agents would entertain the same beliefs. In the standard Bayesian model, this assumption is incarnated in the attribution of the same prior probability to all agents, and it is referred to as the "common prior assumption." Differences in beliefs cannot be commonly known, as

proved by ref. 25 in the celebrated "agreeing to disagree" result.

The common prior assumption has been the subject of heated debates (see refs. 26 and 27, as well as ref. 28 in the context of ref. 29). We believe that studying belief formation processes might shed some light on the reasonability of this assumption. Specifically, when adopting a small-worlds view, positive learning results (such as *Proposition 1*) can identify economic setups where beliefs are likely to be in agreement. By contrast, negative results (such as *Proposition 2*) point to problems where agreement is less likely to be the case.

The literature on polarization asks why agents can become further entrenched in their world views, after observing the same information. In ref. 30 disagreement is possible because agents have different priors and use their current beliefs to interpret ambiguous signals. In ref. 31 disagreement can occur when agents observe imperfect private information about an ancillary variable that affects the interpretation of evidence about the proposition of interest. This paper can be viewed as contributing to this literature, suggesting that, if the ε-EOSF is not unique or is hard to compute, agents might focus on different variables and interpret new observations differently.

ACKNOWLEDGMENTS. We thank Yotam Alexander, Thibault Gajdos, Ed Green, Abhimanyu Gupta, Offer Lieberman, Yishay Mansour, David Schmeidler, and Fan Wang for comments and references. I.G. gratefully acknowledges Israel Science Foundation Grant 704/15.

1. Nadaraya EA (1964) On estimating regression. *Theor Probab Appl* 9:141–142.
2. Watson GS (1964) Smooth regression analysis. *Sankhyā Indian J Stat Ser A* 26:359–372.
3. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297.
4. Vapnik V (2000) *The Nature of Statistical Learning Theory* (Springer Science & Business Media, New York).
5. Medin DL, Schaffer MM (1978) Context theory of classification learning. *Psychol Rev* 85:207–238.
6. Nosofsky RM (1984) Choice, similarity, and the context theory of classification. *J Exp Psychol Learn Mem Cogn* 10:104–114.
7. Jäkel F, Schölkopf B, Wichmann FA (2009) Does cognitive science need kernels? *Trends Cogn Sci* 13:381–388.
8. Nosofsky RM (2011) The generalized context model: An exemplar model of classification. *Formal Approaches in Categorization*, eds Pothos EM, Wills AJ (Cambridge Univ Press, Cambridge, UK), pp 18–39.
9. Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science* 237:1317–1323.
10. Nosofsky RM (1986) Attention, similarity, and the identification–categorization relationship. *J Exp Psychol Gen* 115:39–57.
11. Nosofsky RM (1991) Tests of an exemplar model for relating perceptual classification and recognition memory. *J Exp Psychol Hum Perception Perform* 17: 3–27.
12. Gayer G, Gilboa I, Lieberman O (2007) Rule-based and case-based reasoning in housing prices. *BE J Theor Econ* 7:1935–1704.
13. Lieberman O (2010) Asymptotic theory for empirical similarity models. *Econometric Theor* 26:1032–1059.
14. Lieberman O (2012) A similarity-based approach to time-varying coefficient nonstationary autoregression. *J Time Ser Anal* 33:484–502.
15. Lieberman O, Phillips PCB (2014) Norming rates and limit theory for some time-varying coefficient autoregressions. *J Time Ser Anal* 35:592–623.
16. Lieberman O, Phillips PCB (2017) A multivariate stochastic unit root model with an application to derivative pricing. *J Econom* 196:99–110.
17. Hardlë W, Marron JS (1985) Optimal bandwidth selection in nonparametric regression function estimation. *Ann Stat* 13:1465–1481.
18. Billot A, Gilboa I, Samet D, Schmeidler D (2005) Probabilities as similarity-weighted frequencies. *Econometrica* 73:1125–1136.
19. Gilboa I, Lieberman O, Schmeidler D (2006) Empirical similarity. *Rev Econ Stat* 88:433–444.
20. Billot A, Gilboa I, Schmeidler D (2008) Axiomatization of an exponential similarity function. *Math Soc Sci* 55:107–115.
21. Eilat R (2006) Computational tractability of searching for optimal regularities. MA thesis (Tel Aviv University, Tel Aviv).
22. Aragones E, Gilboa I, Postlewaite A, Schmeidler D (2005) Fact-free learning. *Am Econ Rev* 95:1355–1368.
23. Vapnik V (1998) *Statistical Learning Theory* (Wiley, New York).
24. Savage LJ (1954) *The Foundations of Statistics* (Wiley, New York); 2nd Revised Ed (1972) (Dover, New York).
25. Aumann RJ (1976) Agreeing to disagree. *Ann Stat* 4:1236–1239.
26. Morris S (1995) The common prior assumption in economic theory. *Econ Philos* 11:227–253.
27. Gul F (1998) A comment on Aumann's Bayesian view. *Econometrica* 66:923–927.
28. Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. *Econometrica* 6:1391–1402.
29. Aumann RJ (1987) Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55:1–18.
30. Fryer RG, Harms P, Jackson MO (2018) Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *J Eur Econ Assoc*, 10.1093/jeea/jvy025.
31. Benoît J-P, Dubra J (April 15, 2019) When do populations polarize? An explanation. *Int Econ Rev*, 10.1111/iere.12400.