Second-Order Induction in Prediction Problems*

Rossella Argenziano†, Itzhak Gilboa‡

Abstract

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, non-uniqueness 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.

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 [1] and [2] as well as support vector machines in [3] and [4]). Surprisingly, a very similar formula appeared in the psychological literature in the context of the Generalized Context Model (GCM) (see [5] and [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

*We thank Yotam Alexander, Thibault Gajdos, Ed Green, Offer Lieberman, Yishay Mansour, and David Schmeidler for comments and references. Gilboa gratefully acknowledges ISF Grant 704/15.

†Department of Economics, University of Essex, Wivenhoe Park, CO4 3SQ, Colchester, U.K.

‡Economics and Decision Sciences Department, HEC Paris, 1 rue de la Liberation, 78351 Jouy-en-Josas cedex FRANCE; Berglas School of Economics, Tel-Aviv University, Tel Aviv 6997801, ISRAEL.

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 [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 [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 [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 (see [6], [10], and [11]).

Inspired by these results on classification tasks, we present a model of prediction based on *two levels of learning*.[1] 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_i's. 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 [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)*.[2] Because this process deals with learning how first-order induction should be performed, it will be dubbed *second-order induction*. In

[1] We refer here to the learning needed in order to form prior beliefs, and not to Bayesian learning that such beliefs may later be used for.

[2] [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. [13], [14], [15] and [16] analyzed the asymptotic properties of such estimators. The asymptotic results in this literature assume a given DGP (typically, using a formula such as (1.1), with a noise variable, as the "true" statistical model), whereas our results are more agnostic about the underlying DGP.

statistical terms, this is akin to finding the optimal kernel to estimate y_p. (See [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 NPC (Theorem 3).

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, an agent tries to estimate the probability of water boiling. In the second, 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 is a relatively limited number of variables to take into account, such as temperature, pressure, and a few other experimental conditions. In this type of problems 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, neither by experimentation nor 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 examples 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.

1 Model

1.1 Case-Based Beliefs

The basic problem we deal with is predicting a value of a variable $y \in \mathbb{R}$ based on other variables $x^1, ..., x^m \in \mathbb{R}$. 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 x's, 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, $x_p = \left(x_p^1, ..., x_p^m\right)$, and the corresponding ones for each past observation, $x_i = \left(x_i^1, ..., x_i^m\right)$, 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 $j \in M \equiv \{1, ..., m\}$ for $m \geq 0$. When no confusion is likely to arise, we will refer to the predictor as a "variable" and also refer to the index as designating the variable. The predictors $x \equiv \left(x^1, ..., x^m\right)$ 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"), $x_i = \left(x_i^1, ..., x_i^m\right) \in \mathbb{R}^m$, 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

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)}. \tag{1.1}$$

In case $s(x_i, x_p) = 0$ for all $i \leq n$, we set $\overline{y}_p^s = y_0$ for an arbitrary value $y_0 \in \mathbb{R}$.[3] 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 (1.1) was axiomatized in [18] and in [19].

We use the similarity function [4]

$$s^w(x, x') = \exp\left(-\sum_{j=1}^{m} w^j \left(x^j - x'^j\right)^2\right) \tag{1.2}$$

with $w_j \geq 0$.

Similarity functions that are negative exponentials of norms on the Euclidean space were axiomatized by [20]. [19] and [12] specified the norm to be a weighted Euclidean distance. We will use the extended non-negative 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 j-th summand

[3] We choose some value y_0 only to make the expression $\bar{y}_p^s$ well-defined. Its choice will have no effect on our analysis.

[4] Our results hold for other similarity functions as well. First, the distance it is based on may be based on any semi-norm 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 semi-norms because some w^j's 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 ∞.

in (1.2) 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 (1.2) where the summation is taken over all j's such that $w^j < \infty$.

1.2 Empirically Optimal Similarity Function

How do individuals select a similarity function? Evidence in [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.[5]

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 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 to 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[6]

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

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 $supp(w) \equiv \{ j \leq m \mid w^j > 0 \}$

[5] Notice that the axiomatic derivations mentioned above ([18],[19],[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 – relatively infrequently. Thus the axiomatic derivations hold approximately, and the appropriate similarly function is learnt over longer time spans.

[6] 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.

and $c > 0$. We will also use $supp(A)$ to denote the set of supports of all 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 $0 < c < c_0$,

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

(Observation 1 is proved in the *S.I.*) 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 will therefore be interested in vectors w that obtain the lowest $AMSE$ approximately. More precisely, we define *ε-empirically optimal similarity function* as follows:

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

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

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 will 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.

2 Results

2.1 Almost-Uniqueness

In this section we provide three results. Their proofs are contained in the *S.I.* 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 $y = f\left(x\right)$. 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 is in the bounded interval $[-K, K]$ for $K > 0$. Let g be the joint density of x, with $g\left(x\right) \geq \eta > 0$ 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 $y_i = f\left(x_i\right)$.[7] Refer to this data generating process as

[7] A similar result would hold if we allow y_i to be distributed around $f\left(x_i\right)$ with an i.i.d. error term.

(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 indices 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 $W \geq 0$ 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*

$$P\left(MSE\left(w\right) < \nu\right) \geq 1 - \xi.$$

where the probability $P = P\left(n, m, g, f\right)$ is the measure induced by the process described above.
Conversely, if $j \in I(f)$, then for every $W \geq 0$ and $\xi > 0$ there exist $\nu > 0$ and N such that, if $w^j \leq W$, then, for every $n \geq N$,

$$P\left(MSE\left(w\right) > \nu\right) \geq 1 - \xi.$$

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

$$P\left(w \in \varepsilon\text{-}\arg\min AMSE \implies supp(w) = I(f)\right) \geq 1 - \xi$$

The proposition 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.

The proposition 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 (see [1], [2], and [17]) although we are unaware of a statement of a result that directly implies this one. On the other hand, the second part of the proposition 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 the weight vectors in ε-$\arg\min AMSE$ share the same support, namely the set of informative variables.

Denoting a "ball" of ∞ as $N_W\left(\infty\right) = \left\{ w \in [0, \infty)^m \;\middle|\; w^j \geq W \quad \forall j \leq m \right\}$, the first part of the proposition states that, given (a small) $\nu > 0$, there exists (a large) W such that any point in $N_W\left(\infty\right)$ 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 $N_W\left(\infty\right)$.

These first two parts of the proposition 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, the proposition 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 the 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 degrees, 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 x's are not constrained to be independent.[8]

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 non-unique 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 non-empty 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 simi-

[8] The assumption of independence of the y_i's is only used to guarantee that each observation y_i has sufficiently close other observations, and it can therefore be significantly relaxed.

larity function. But the following proposition suggests that, assuming a process as discussed here, the non-uniqueness 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 the variables in $J \cup J'$. Hence, agents who seek an ε-EOSF to explain the data may believe either that J as 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. [9]

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 of the observed x_i. Thus, there is something to be learnt. 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;[10] but the number of variables grows relative to n. 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

[9] The proof shows that these sets can also be disjoint .

[10] Holding n fixed, a perfect fit for the y_i's will not be obtained even if m grows to infinity.

any large enough set of variables, there will be disjoint sets that provide near-perfect fit, and thus the ε-EOSF will be non-unique in a way that we think of as "fundamental".

2.2 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. 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 non-negative numbers w such that $AMSE(w, c) \leq R$?

And we can state

Theorem 3. *ε-EOSF is NPC.*

Theorem (3) states that Problem (2.2) 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, that 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 (3) 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.[11]

2.3 Second-Order Induction and Learnability

Our analysis can be viewed as adding to a large literature on what can and what cannot be learnt. 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 set-ups:

(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 can't control the observed x's;

(iii) There is no bounded set of variables x such that y_i depends only on x_i, independently of past values.

[11] See also [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 [22]).

Set-up (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 set-up (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 (see [23]) suggests the VC dimension of the set of possible functional relationships as a litmus test for the classes of functions that can be learnt and those that cannot. Finally, there are problems that are closer to set-up (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 types (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 set-up (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 (see [1] and [2]). Indeed, Proposition 1 is in the spirit of [17] in showing that, as n grows, kernel estimation with optimal kernel parameters leads to good predictions. However, [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 learnt 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 [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 3) 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.

3 Discussion

3.1 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 OLS (ordinary least square) 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 NP-Hard (see [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 only predicted 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 x's.[12]

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.

3.2 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 [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.

[12] This is also the reason that Proposition 2 required a large n before demanding that m be large relative to n.

3.3 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 [25] in the celebrated "agreeing to disagree" result.

The Common Prior Assumption has been the subject of heated debates (see [26], [27], as well as [28] in the context of [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 set-ups 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 [30] disagreement is possible because agents have different priors and use their current beliefs to interpret ambiguous signals. In [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 isn't unique or is hard to compute, agents might focus on different variables and interpret new observations differently.

References

[1] Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability & Its Applications* **9**, 141–142.

[2] Watson, G. S. (1964) Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* pp. 359–372.

[3] Cortes, C & Vapnik, V. (1995) Support-vector networks. *Machine learning* **20**, 273–297.

[4] Vapnik, V. (2013) *The nature of statistical learning theory.* (Springer science & business media).

[5] Medin, D. L & Schaffer, M. M. (1978) Context theory of classification learning. *Psychological review* **85**, 207.

[6] Nosofsky, R. M. (1984) Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition* **10**, 104.

[7] Jäkel, F, Schölkopf, B, & Wichmann, F. A. (2009) Does cognitive science need kernels? *Trends in cognitive sciences* **13**, 381–388.

[8] Nosofsky, R. M. (2011) The generalized context model: An exemplar model of classification. *Formal approaches in categorization* pp. 18–39.

[9] Shepard, R. N. (1987) Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323.

[10] Nosofsky, R. M. (1986) Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General* **115**, 39.

[11] Nosofsky, R. M. (1991) Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of experimental psychology: human perception and performance* **17**, 3.

[12] Gayer, G, Gilboa, I, & Lieberman, O. (2007) Rule-based and case-based reasoning in housing prices. *The BE Journal of Theoretical Economics* **7**.

[13] Lieberman, O. (2010) Asymptotic theory for empirical similarity models. *Econometric Theory* **26**, 1032–1059.

[14] Lieberman, O. (2012) A similarity-based approach to time-varying coefficient non-stationary autoregression. *Journal of Time Series Analysis* **33**, 484–502.

[15] Lieberman, O & Phillips, P. C. (2014) Norming rates and limit theory for some time-varying coefficient autoregressions. *Journal of Time Series Analysis* **35**, 592–623.

[16] Lieberman, O & Phillips, P. C. (2017) A multivariate stochastic unit root model with an application to derivative pricing. *Journal of Econometrics* **196**, 99–110.

[17] Hardlë, W & Marron, J. S. (1985) Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics* **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. *The Review of Economics and Statistics* **88**, 433–444.

[20] Billot, A, Gilboa, I, & Schmeidler, D. (2008) Axiomatization of an exponential similarity function. *Mathematical Social Sciences* **55**, 107–115.

[21] Eilat, R. (2007) Computational tractability of searching for optimal regularities. *working paper*.

[22] Aragones, E, Gilboa, I, Postlewaite, A, & Schmeidler, D. (2005) Fact-free learning. *American Economic Review* **95**, 1355–1368.

[23] Vapnik, V. (1998) *Statistical learning theory. 1998.* (Wiley, New York) Vol. 3.

[24] Savage, L. J. (1954) *The foundations of statistics.* (New York: John Wiley and Sons. (Second addition in 1972, Dover)).

[25] Aumann, R. J. (1976) Agreeing to disagree. *The annals of statistics* pp. 1236–1239.

[26] Morris, S. (1995) The common prior assumption in economic theory. *Economics & Philosophy* **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* pp. 1391–1402.

[29] Aumann, R. J. (1987) Correlated equilibrium as an expression of bayesian rationality. *Econometrica* pp. 1–18.

[30] Fryer, Roland G.and Harms, P & Jackson, M. O. (2018) Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association.* https://doi.org/10.1093/jeea/jvy033.

[31] Benoit, J.-P & Dubra, J. (2018) When do populations polarize? an explanation. *working paper.*

[32] Garey, M. R & Johnson, D. S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness.* (San-Francisco, CA:W. Freeman and Co.).

Supporting Information: Proofs

Proof of Observation 1

Assume that $m = 1$, $n = 4$ and

i	x_i	y_i
1	0	0
2	1	0
3	3	1
4	4	1

In this example observations $1, 2$ are closer to each other than each is to any of observations $3, 4$ and vice versa. (That is, $|x_i - x_j| = 1$ for $i = 1, j = 2$ as well as for $i = 3, j = 4$, but $|x_i - x_j| \geq 2$ for $i \leq 2 < j$.) Moreover the values of y are the same for the "close" observations and different for "distant" ones. (That is, $y_i = y_j$ for $i = 1, j = 2$ as well as for $i = 3, j = 4$, but $|y_i - y_j| = 1$ for $i \leq 2 < j$.) If we choose a finite w, the estimated value for each i, $\bar{y}_i^{s_w}$, is a weighted average of the two distant observations and the single close one. In particular, for every $w < \infty$ we have $MSE(w) > 0$.

Observe that $w = w^1 = \infty$ doesn't provide a perfect fit either: if we set $w = w^1 = \infty$, each observation i is considered to be dissimilar to any other, and its y value is estimated to be the default value, $\bar{y}_i^{s_w} = y_0$. Regardless of the (arbitrary) choice of y_0, the MSE is bounded below by that obtained for $y = 0.5$ (which is the average y in the entire database). Thus, $MSE(\infty) \geq 0.25$.

Thus, $MSE(w) > 0$ for all $w \in [0, \infty]$. However, as $w \to \infty$ (but $w < \infty$), for each i the weight of the observation that is closest to i converges to 1 (and the weights of the distant ones – to zero), so that $\bar{y}_i^{s_w} \to y_i$. Hence, $MSE(w) \to_{w \to \infty} 0$. We thus conclude that $\inf_{w \in [0,\infty]} MSE(w) = 0$ but that there is no w that minimizes the MSE.

The same argument applied to the $AMSE(w, c)$ for any $c < c_0$ if we set $c_0 = 0.25$. □

Proof of Proposition 1

We first wish to show that arbitrarily low values of the MSE can be obtained with probability that is arbitrarily close to 1, provided the weights w^j are all large enough. Let there be given $\nu > 0$ and $\xi > 0$. We wish to find N and W such that for every $n \geq N$, and every vector w such that $w^j \geq W$ but $w^j < \infty$ $(\forall j \leq m)$ we have

$$P(MSE(w) < \nu) \geq 1 - \xi.$$

Observe that a single j for which $w^j = \infty$ suffices to set the MSE at least as high as the variance of (y_i), as, with probability 1, each observation will be the unique one with the specific value of x^j.

We now define "proximity" of the x values that would guarantee "proximity" of the y values. Suppose that the latter is defined by $\nu/2$. As the function f is continuous on a compact set, it is uniformly continuous. Hence, there exists $\theta > 0$ such that, for any x, x' that satisfy $\|x - x'\| < \theta$ we have $[f(x) - f(x')]^2 < \nu/2$. Let us divide the set X into $(4K\sqrt{m}/\theta)^m$ equi-volume cubes, each with

an edge of length $\frac{\theta}{2\sqrt{m}}$. Two points x, x' that belong to the same cube differ by at most $\frac{\theta}{2\sqrt{m}}$ in each coordinate and thus satisfy $\|x - x'\| < \theta/2$. Let us now choose N_1 such that, with probability of at least $(1 - \xi/2)$, each such cube contains at least two observations x_i $(i \leq N_1)$. This guarantees that, when observation i is taken out of the sample, there is another observation i' (in the same cube), with $[y_{i'} - f(x_i)]^2 < \nu/2$.

Next, we wish to bound the probability mass of each cube (defined by g). The volume of a cube is $\left(\frac{\theta}{2\sqrt{m}}\right)^m$ and the density function is bounded from below by η. Thus, the proportion of observations in the cube (out of all the n observations) converges (as $n \to \infty$) to a number that is bounded from below by $\zeta \equiv \eta \left(\frac{\theta}{2\sqrt{m}}\right)^m > 0$. Choose $N \geq N_1$ such that, with probability of at least $(1 - \xi/2)$, for each $n \geq N$ the proportion of the observations in the cube is at least $\zeta/2$. Note that this is a positive number which is independent of n.

We can now turn to choose W. For each i, the proportion of observations x_k with $[f(x_i) - f(x_k)]^2 > \nu$ is bounded above by $(1 - \zeta)$. Choose w such that $w^j = W$. Observe that, as $W \to \infty$,

$$\frac{\sum_{k \neq i, [f(x_i) - f(x_k)]^2 > \nu} s(x_i, x_k)}{\sum_{k \neq i, [f(x_i) - f(x_k)]^2 \leq \nu} s(x_i, x_k)} \to 0$$

and this convergence is uniform in n (as the definition of ζ is independent of n). Thus a sufficiently high W can be found so that, for all $n \geq N$, $MSE(w_0) < \nu$ with probability $(1 - \xi)$ or higher.

Next we prove the second part of the proposition. Assume that x^j is informative, so that there exist x, x' such that $x^l = x'^l$ for all $l \neq j$ but $f(x) - f(x') = \delta > 0$. Assume that, for some $W < \infty$, $w^j \leq W$. Similar arguments to those above yield an lower bound $\nu > 0$ such that, for large n, with very high probability, $MSE(w) > \nu$: points around x will have estimated y values that are affected by points around x', and the weight of these will not converge to zero (it is bounded from below by e^{-W}).

Finally, we wish to show that one can have a low enough cost c_0 such that all the vectors in ε-$\arg\min AMSE$ would use the informative variables, as well as a low enough ε so that they would not use the uninformative variables. This would mean that for appropriately chosen c_0 and ε, the supports of all vectors in ε-$\arg\min AMSE$ have to coincide with $I(f)$. Let there be given $\xi > 0$. For each $j \in I(f)$ we can use the second part of the proposition (corresponding to $W = 0$) to find $\nu_j > 0$ and N_j such that, for every $n > N_j$, with probability of at least $(1 - \xi/2m)$, $w^j = 0$ implies $MSE(w) > \nu_j$. Define $N_j = 0$ for $j \notin I(f)$.

Choose $c_0 = min_j(\nu_j)/2(m+1)$ and let $c < c_0$. Using the first part of the proposition, let N_0 and W_0 be such that, for all $n \geq N_0$, with probability of at least $(1 - \xi/2)$, $MSE(w_0) < c$ for w_0 defined by $w_0^l = W_0$ for all l. Consider $N = max(N_l)_{l \geq 0}$. For every $n \geq N$, with probability of at least $(1 - \xi)$ we have that (i) there are w with $MSE(w) < c$; (ii) for these w's,

$AMSE(w) < (m+1)c$; (iii) for any vector w whose support does not include $j \in I(f)$, $AMSE(w) > \nu_j > (m+1)c$. This means that for every w with $AMSE(w) < (m+1)c$, we must have $I(f) \subset supp(w)$. Thus, considering near-minimizers of the $AMSE$ we will only find vectors that use all the informative variables. On the other hand, we wish to show that in the (high-probability) event considered above, variables that are not informative will not be used. Observe that $\varepsilon < c/2$ is small enough so that for every $w \in \varepsilon\text{-arg}\min AMSE$, $w^j = 0$ for every $j \notin I(f)$ (as the inclusion of such a variable in the support of w would incur a cost that is by itself enough to make the $AMSE$ of the vector larger than the argmin by more than $\varepsilon < c/2$).

□

Proof of Proposition 2:

Non-uniqueness is obtained by showing that, with a high probability there will be two variables, each of which can provide an almost perfect fit on its own. To this end, we first need to make sure that each observation y_i has a close enough y_k. For this reason the result only holds for a relatively large n (making sure that, with a high probability, no y_i is "isolated"), and then, given such an n, for a large enough number of predictors, $M(n)$, so that we should think of this case as $m >> n >> 1$.

We now turn to prove the result formally. Let there be given $c > 0$. Choose $\bar{\varepsilon} = c/3$. We wish it to be the case that if $MSE(w) \leq \varepsilon$ with $\#supp(w) = 1$, then $w \in \varepsilon\text{-arg}\min AMSE$, but for no $w \in \varepsilon\text{-arg}\min AMSE$ is it the case that $\#supp(w) > 1$. Clearly, the choice $\bar{\varepsilon} = c/3$ guarantees that for every $\varepsilon \in (0, \bar{\varepsilon})$, the second part of the claim holds: if a vector w satisfies $MSE(w) \leq \varepsilon$, no further reduction in the MSE can justify the cost of additional variables, which is at least c. Conversely, because $c < v/2$ (the variance of y), a single variable j that obtains a near-zero MSE would have a lower $AMSE$ than the empty set.

Let there now be given $\varepsilon \in (0, \bar{\varepsilon})$ and every $\delta > 0$. We need to find N and, for every $n \geq N$, $M(n)$, such that for every $n \geq N$ and $m \geq M(n)$,

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

Let N be large enough so that, with probability $(1 - \delta/2)$, for all $n \geq N$,

$$\max_i \min_{k \neq i} [y_i - y_j] < \varepsilon/2.$$

(To see that such an n can be found, one may divide the $[-K, K]$ interval of values to intervals of length $\varepsilon/2$ and choose N to be large enough so that, with the desired probability, there are at least two observations in each such interval.)

Given such $n \geq N$ and the realizations of $(y_i)_{i \leq n}$, consider the realizations of x^j. Assume that, for some j, it so happens that $\left|x_i^j - y_i\right| < \varepsilon/4$ for all $i \leq n$. In this case, by setting w^j to be sufficiently high, and $w^l = 0$ for $l \neq j$, one would obtain $MSE(w) \leq \varepsilon$ and $AMSE(w, c) \leq \varepsilon + c$.[13] For each j,

[13]The fact that x_i^j is close to y_i is immaterial, of course, as the variables x_i^j are not used to predict y_i directly, but only to identify the y_k that would. If x_i^j is close to some monotone function of y_i the same argument would apply.

however, the probability that this will be the case is bounded below by some $\xi > 0$, independent of n and j. Let $M_1(n)$ be a number such that, for any $m \geq M_1(n)$, the probability that at least one such j satisfies $\left|x_i^j - y_i\right| < \varepsilon/4$ is $(1 - \delta/4)$, and let $M(n) > M_1(n)$ be a number such that, for any $m \geq M(n)$, the probability that at least one more such $j' > j$ satisfies $\left|x_i^{j'} - y_i\right| < \varepsilon/4$ is $(1 - \delta/8)$.

Thus, for every $n \geq N$, and every $m \geq M(n)$, with probability $1 - \delta$ there are two vectors, w^j with support $\{j\}$ and $w^{j'}$ with support $\{j'\}$, each of which obtaining $MSE(w) \leq \varepsilon$ and thus, both belonging to ε-arg min $AMSE$. To see that in this case the *supp*(ε-arg min $AMSE$) is not closed under union, it suffices to note that no w with support greater than a singleton, nor a w with an empty support (that is, $w \equiv 0$) can be in the ε-arg min $AMSE$. □

Proof of Theorem 1

We first verify that the problem is in NP. Given a database and a vector of extended rational weights $w^j \in [0, \infty]$, the calculation of the $AMSE$ takes $O\left(n^2 m\right)$ steps. Specifically, the calculation of the similarity function $s(x, x')$ is done by first checking whether there exists a j such that $w^j = \infty$ and $x^j \neq x'^j$ (in which case $s(x, x')$ is set to 0), and, if not – by ignoring the j's for which $w^j = \infty$.

The proof is by reduction of the SET-COVER problem to EMPIRICAL-SIMILARITY. The former, which is known to be NPC (see [32]), is defined as

Problem. SET-COVER: Given a set P, $r \geq 1$ subsets thereof, $T_1, ..., T_r \subseteq P$, and an integer k $(1 \leq k \leq r)$, are there k of the subsets that cover P? (That is, are there indices $1 \leq i_1 \leq i_2 \leq ... \leq i_k \leq r$ such that $\cup_{j \leq k} T_{i_j} = P$?)

Given an instance of SET-COVER, we construct, in polynomial time, an instance of EMPIRICAL-SIMILARITY such that the former has a set cover iff the latter has a similarity function that obtains the desired AMSE. Let there be given P, $r \geq 1$ subsets thereof, $T_1, ..., T_r \subseteq P$, and an integer k. Assume without loss of generality that $P = \{1, ..., p\}$, that $\cup_{i \leq r} T_i = P$, and that $z_{uv} \in \{0, 1\}$ is the incidence matrix of the subsets, that is, that for $u \leq p$ and $v \leq r$, $z_{uv} = 1$ iff $u \in T_v$.

Let $n = 2(p + 1)$ and $m = r$. Define the database $B = ((x_i, y_i))_{i \leq n}$ as follows. (In the database each observation is repeated twice to avoid bins of size 1.)

For $u \leq p$ define two observations, $i = 2u - 1, 2u$ by

$$x_i^j = z_{uj} \qquad y_i = 1$$

and add two more observations, $i = 2p + 1, 2p + 2$ defined by

$$x_i^j = 0 \qquad y_i = 0.$$

Next, choose c to be such that $0 < c < \frac{1}{mn^3}$, say, $c = \left(mn^3\right)^{-1}/2$ and $R = kc$.[14] This construction can obviously be done in polynomial time.

We claim that there exists a vector w with $AMSE(w, c) \leq R$ iff a cover of size k exists for the given instance of SET-COVER.[15] For the "if" part, assume that such a cover exists, corresponding to $J \subseteq M$. Setting the weights

$$w^j = \begin{cases} \infty & j \in J \\ 0 & j \notin J \end{cases}$$

one obtains $AMSE(w, c) \leq R$.

Conversely, for the "only if" part, assume that a vector of rational weights $w = \left(w^j\right)_j$ ($w^j \in [0, \infty]$) obtains $AMSE(w, c) \leq R$. Let $J \subseteq M$ be the set of indices of predictors that have a positive w^j (∞ included). By the definition of R (as equal to ck), it has to be the case that $|J| \leq k$. We argue that J defines a cover (that is, that $\{T_v\}_{v \in J}$ is a cover of P).

Observe that, if we knew that $|J| = k$, the inequality

$$AMSE(w, c) = MSE(w) + c|J| \leq R = ck$$

could only hold if $MSE(w) = 0$, from which it would follow that w provides a perfect fit. In particular, for every $i \leq 2p$ there exists $j \in J$ such that $x_i^j \neq x_{2p+1}^j$ that is, $x_i^j = 1$, and J defines a cover of P.

However, it is still possible that $|J| < k$ and $0 < MSE(w) \leq c(k - |J|)$. Yet, even in this case, J defines a cover. To see this, assume that this is not the case. Then there exists $i \leq 2p$ such that for all j, either $w^j = 0$ ($j \notin J$) or $x_i^j = 0 = x_{2p+1}^j$. This means that $s(x_i, x_{2p+1}) = s(x_i, x_{2p+2}) = 1$. In particular, $y_{2p+1} = y_{2p+2} = 0$ take part (with positive weights) in the computation of $\overline{y}_i^{s_w}$ and we have $\overline{y}_i^{s_w} < 1 = y_i$. The cases $2p+1, 2p+2$ obtain maximal similarity to i ($s(x_i, x_{2p+1}) = s(x_i, x_{2p+2}) = 1$), because $x_{2p+1}^j = x_{2p+2}^j = x_i^j (= 0)$ for all j with $w^j > 0$. (It is possible that for other observations $l \leq 2p$ we have $s(x_i, x_{2p+1}) \in (0, 1)$,but the weights of these observations are evidently smaller than that of $2p+1, 2p+2$.) Thus we obtain that the error $|\overline{y}_i^{s_w} - y_i|$ must be at least $\frac{1}{n}$, from which $SSE(w) \geq \frac{1}{n^2}$ and $MSE(w) \geq \frac{1}{n^3}$ follow. This implies $AMSE(w, c) > R$ and concludes the proof. □

[14] As will be clarified shortly, the power of n in the constant c reflects the choice of the quadratic loss function. Different loss functions would require a corresponding cost c. For example, for an absolute value $c = \left(mn^2\right)^{-1}/2$ would suffice.

[15] This proof uses values of x and of y that are in $\{0, 1\}$. However, if we consider the same problem in which the input is restricted to be positive-length ranges of the variables, one can prove a similar result with sufficiently small ranges and a value of R that is accordingly adjusted.