Second-Order Induction and the Importance of Precedents[*]

Rossella Argenziano[†] and Itzhak Gilboa[‡]

June 2017

Abstract

Agents make predictions based on similar past cases. The notion of similarity is itself learnt from experience by "second-order induction": past cases inform agents about also the relative importance of various attributes in judging similarity. Second-order induction can explain the importance of precedents: a precedent doesn't change only the empirical frequencies, but also the way similarity is judged. In particular, the model can explain why reputation is harder to re-establish, after having been lost, than to establish a priori. However, there may be multiple "optimal" similarity functions for explaining past data. Moreover, the computation of the optimal similarity function is NP-Hard. As a result, rational agents who have access to the same observations may still entertain different probabilistic beliefs.

1 Introduction

Economic theory tends to assume that rational agents entertain probabilistic beliefs, which may be subjective if probabilities are not given. This assumption is supported by formidable axiomatic derivations by Ramsey, (1926a,b), de Finetti, (1931,1937), Savage, (1954), and Anscombe-Aumann

[*] Gilboa gratefully acknowledges ISF Grant 704/15.

[†] Department of Economics, University of Essex. r_argenziano@essex.ac.uk

[‡] HEC, Paris-Saclay, and Tel-Aviv University. tzachigilboa@gmail.com

(1963). These, however, deal with form rather than content: they can be used to convince decision makers that probability is the way to model beliefs, but they do not specify how these beliefs are to be generated. Indeed, rationality is typically also taken to impose some constraints on the formation of beliefs, such as taking into account past observations. In particular, if objective probabilities happen to exist, rational agents are expected to adopt them as their subjective beliefs. For example, observing repeated i.i.d. tosses of a coin, one expects beliefs about future observations to be close to the observed empirical frequencies.

Empirical frequencies lead to objective probabilities in many economic setups. An agent who considers buying an insurance policy against car theft may access a large database of practically identical cases, and infer what her probabilities should be from the relative frequency of thefts in the database. But there are also many economic problems in which past cases are given, yet they are not entirely identical. In judging the likelihood of an imminent financial crisis or a revolution, one would be silly to ignore past cases. At the same time, one may not assume that all past observations are results of an i.i.d. process. Even if one were to ignore possible causal relationships between past cases and future ones, each case has sufficiently many relevant details to make it unique. Hence, the calculation of empirical frequencies is not as straightforward as in the example of a car theft, let alone in examples of chance games such as the toss of a coin.

We assume that in such problems agents use past data to estimate probabilities, but also use their judgment to determine which cases are more similar, and therefore more relevant for the problem at hand. Using similarity-weighted empirical frequencies is an intuitive idea that appeared both in statistics (as kernel-based probabilities) and in psychology ("exemplar learning"). Thus, it is a model of belief formation that can be interpreted both normatively and descriptively, and it can also be derived axiomatically (Billot, Gilboa, Samet, and Schmeidler, 2005, Gilboa, Lieberman, and Schmeidler, [GLS] 2006). However, such a process begs the question: where would the similarity function come from?

The notion of "second-order induction" suggests that the similarity function should also be learnt from past cases. Using a given similarity function to learn from past cases about future ones can be referred to as "first-order induction"; learning the similarity function, that is, learning how this first-order induction should be done is dubbed "second-order induction". For example, suppose that one has to predict the quality of a car. Children may make similarity judgments based on perceptual features, such as color. Adults will typically realize that the make of the car is a more important feature than its color. Asked why, a person might say, "Experience has shown that cars with the same color may be very different in quality, and vice versa – the exact same car can be found in very different colors. In short, over the years I learned that color isn't an important feature for judging quality."

We model this type of reasoning by looking for a similarity function that provides a good fit for the data. That is, using a leave-one-out cross-validation technique, we look for a function, within a given class, that would have provided the best predictions were it used to predict each point in the database based on the other ones. We refer to it as "the empirical similarity". We suggest this form of learning as an obviously-idealized model of the way people naturally learn which variables are important for similarity judgments. This learning process has been suggested and analyzed in GLS (2006) (in a slightly different model) as a statistical technique. Indeed, it can also be interpreted normatively, as a way of performing non-parametric statistics with a kernel function that is estimated from the data. However, in this paper our focus is descriptive, and we use the model to describe human reasoning.

The paper studies the notion of empirical similarity in a binary model and explores some of its economic consequences. We first point out that the empirical similarity function need not take into account all the variables available. For reasons that have to do both with the curse of dimensionality and with overfitting, one may prefer to use a relatively small set of the variables to a superset thereof. We provide conditions under which it is worthwhile to add a variable to the determinants of the similarity function and discuss their applicability to economic set-ups. Next, we observe that the empirical similarity need not be unique, and that people who have access to the same database may end up using different variables to obtain the "best" fit. Further, we show that finding the best similarity function is a computationally complex (NP-Hard) problem. Thus, even if the empirical similarity is unique, it does not immediately follow that all agents can find it.

We use these theoretical results to offer new ways of looking at some economic phenomena. First, we argue that when the empirical similarity is not unique, or hard to compute, one may expect opinions to differ: people may well use different similarity functions, and, as a result, entertain different beliefs. Second, we argue that the empirical similarity model can help explain the importance of precedents in general, and of reputation in particular. We argue that a precedent isn't just an observation that changes empirical frequencies; it also changes the way that empirical frequencies are weighted. As a result, economic agents would be willing to invest considerable resources in establishing or in preventing precedents, above and beyond their impact on relative frequencies.

The following subsection provides a motivating example. Section 2 presents the basic model and the idea of the empirical similarity formally. Section 3 deals with the general questions of monotonicity, uniqueness, and computational complexity of the empirical similarity function. In Section 4 we illustrate some applications of the model. Generalizations to non-binary models are discussed in Section 5. Finally, Section 6 concludes with a general discussion.

1.1 The Election of President Obama

The election of Obama as President of the US in 2008 was a defining event in US history. For the first time, a person who defines himself and is perceived by others as an African-American was elected for the highly coveted office, and this was clearly an important precedence. Whereas in the past African-Americans would have thought that they had no chance of being elected, as there had been no cases of presidents of their race, now there was such a case, and the statistics started looking differently.

It appears, however, that the single case of President Obama changes statistics far beyond its relative frequency, and this remains true even if we weigh cases by their recency. For example, considering only the post-WWII period, the US had 11 presidents before Obama. The effect of his election, however, does not seem to be captured by the difference between 0:11 and 1:12. The precedent set by Obama, we argue, is partly explained by second-order induction: whereas, up to his election, the empirical similarity might include "race" as a relevant variable for similarity judgment, this may no longer be the case after the election. More generally, the effect of a precedent is not only in the empirical frequencies as weighted by a given similarity function, but in changes in this function.

2 Case-Based Beliefs

2.1 The Basic Model

The basic problem we deal with is predicting a value of a variable y based on other variables $x^1, ..., x^m$. 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. This problem is, of course, a standard one in statistics and in machine learning. However, in these fields the goal is basically to find a prediction method that does well according to some criteria. By contrast, our interest is in modeling how people tend to reason about such problems. Luckily, the two questions are not divorced from each other. For example, while linear regression is the basic workhorse of statistics for more than a century, it has also been used as a model of reasoning of economic agents (see Bray, 1982). Similarly, non-parametric statistics suggested kernel methods (see Akaike, 1954, Rosenblatt, 1956, Parzen, 1962, and Silverman, 1986) which turned out to be equivalent to models of human reasoning. Specifically, a kernel-weighted average is equivalent to "exemplary learning" in psychology, and various kernel techniques ended up being identical to similarity-based techniques axiomatized in decision theory. (See Gilboa and Schmeidler, 2012.)[1]

Following the literature in psychology and artificial intelligence about exemplary learning and case-based reasoning, as well as the case-based decision and case-based prediction projects of Gilboa and Schmeidler (2001, 2012), we focus here on prediction by rather basic case-based formulae. These are equivalent to kernel methods, but we stick to the terms "cases" and "similarity" – rather than "observations" and "kernel" – in order to emphasize the descriptive interpretation adopted here: our goal is not to find the best prediction technique, but to come up with a reasonable model of the way people think.

We will therefore assume that prediction is made based on a similarity function $s : X \times X \rightarrow \mathbb{R}_+$. 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, $x_i = (x_i^1, ..., x_i^m)$, so that $s(x_i, x_p)$ would measure the degree to which the past case is similar to the present

[1] Gilboa and Schmeidler (2012) discuss some examples where certain mathematical models have been independently developed in statistics (with a normative goal in mind) and in psychology (attempting to model human reasoning). In some cases, known biases of statistical methods (such as kernel estimation) re-appear as psychological biases (see Gayer, 2009).

The concurrence between techniques that are designed by statisticians and the techniques that the human brain seems to be implementing might be encouraging if we think of the optimality of the brain, or discouraging when we focus on the brain's limitations in innovating beyong its own technology.

one. The similarity function should incorporate not only intrinsic similarity judgments, but also judgments of relevance, probability of recall and so forth.[2]

In this paper we focus on a binary model, according to which all the variables – the predictors, $x^1, ..., x^m$, and the predicted, y – as well as the weights of the variables in the similarity function and the similarity function itself take values in $\{0, 1\}$. This is obviously a highly simplified model that is used to convey some basic points.[3] We later discuss the generalization of the model to continuous variables and continuous weights.

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 $x \equiv (x^1, ..., x^m)$ assume values (jointly) in $X \equiv \{0, 1\}^m$ and the predicted variable, y, – in $\{0, 1\}$. 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"), $x_i = (x_i^1, ..., x_i^m) \in X$, and $y_i \in \{0, 1\}$, and $x_p \in X$ is a new data point. The goal is to predict the value of $y_p \in \{0, 1\}$ corresponding to x_p, or, more generally, to estimate its distribution.

Given a function $s : X \times X \to \{0, 1\}$, the probability that $y_p = 1$ is estimated by

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

if $\sum_{i \leq n} s(x_i, x_p) > 0$ and $\overline{y}_p^s = 0.5$ otherwise.

This formula is identical to the kernel-averaging method (where the similarity s plays the role of the kernel function). Gilboa, Lieberman, and Schmeidler (2006) provide axioms on likelihood judgments (conditioned on

[2] Typically, the time at which a case occurred would be part of the variables x, and thus recency can also be incorporated into the similarity function.

[3] Moreover, in our model the similarity relation defined by similarity value of 1 will be an equivalence relation. From a conceptual viewpoint, this may be the first assumption one would like to relax.

databases) that are equivalent to the existence of a function s such that (5) holds for any database B.[4] Because the similarity function in our model only takes values in $\{0,1\}$, it divides the database into observations (x_i, y_i) whose x values are similar (to degree 1) to x_p, and those who are not (that is, similar to degree 0), and estimates the probability that y_p be 1 by the relative empirical frequencies of 1's in the sub-database of similar observations.

Finally, we specify the similarity function as follows: given weights for the variables, $(w^1, ..., w^m) \in X \; (\equiv \{0,1\}^m)$, let

$$s_w(x_i, x_p) = \prod_{\{j | w^j = 1\}} \mathbf{1}_{\{x_i^j = x_p^j\}} \tag{2}$$

Thus, the weights $(w^1, ..., w^m)$ determine which variables are taken into consideration, and the similarity of two vectors is 1 iff they are identical on these variables. Clearly, the relation "having similarity 1" is an equivalence relation.

2.2 Empirical Similarity

Where does the similarity function come from? The various axiomatic results mentioned above state that, under certain conditions on likelihood or probabilistic judgments, such a function exists, but they do not specify which function it is, or which functions are more reasonable for certain applications than others. The notion of second-order induction is designed to capture the idea that the choice of a similarity function is made based on data as well. It is thus suggested that, within a given class of possible functions, $\mathcal{S}$, one choose a function that fits the data best. Finding the weights w such that, when fed into s_w, fit the data best renders the empirical similarity problem parametric: while the prediction of the value of y is done in a non-parametric way (as in kernel estimation), relying on the entire database for

[4] See also Billot, Gilboa, Samet, and Schmeidler (2005) for the similarity-weighted averaging of probability vectors with more than two entries.

each prediction, the estimation of the similarity function itself is reduced to the estimation of m parameters.

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 "leave one out": 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 the probability that $y_i = 1$, 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 sum of squared errors it would have had. A similarity function that minimizes this number is chosen as the "empirical similarity".

Explicitly, let there be given a set of similarity functions $\mathcal{S}$. (In our case, $\mathcal{S} = \{\, s_w \mid w \in X \,\}$.) For $s \in \mathcal{S}$, let

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

if $\sum_{j\neq i} s(x_j, x_i) > 0$ and $\overline{y}_i^s = 0.5$ otherwise. Define the sum of squared errors to be

$$SSE\,(s) = \sum_{i=1}^{n} \left(\overline{y}_i^s - y_i\right)^2$$

It will also be convenient to consider the mean (squared) error, that is,

$$MSE\,(s) = SSE\,(s)\,/n.$$

It will be useful to define, for a set of variables indexed by $J \subseteq M$, the indicator function of J, w_J, that is,

$$w_J^l = \begin{cases} 1 & l \in J \\ 0 & l \notin J \end{cases} .$$

To simplify notation, we will use $SSE\,(J)$ for $SSE\,(s_{w_J})$ and $MSE\,(J)$ for $MSE\,(s_{w_J})$.

The similarity functions we consider divide the database into sub-databases, or "bins", according to the values of the variables in J. Formally, for $J \subseteq M$ and $z \in \{0,1\}^J$, define the *J-z bin* to be the cases in B that correspond to these values. Formally, we will refer to the set of indices, that is,

$$b(J,z) = \left\{ i \leq n \;\middle|\; x_i^j = z^j \quad \forall j \in J \right\}$$

as "the *J-z bin*".

It will also be convenient to define, for $J \subseteq M$, and $z \in \{0,1\}^J$, $j \in M \backslash J$, and $z^j \in \{0,1\}$, the bin obtained from adding the value z^j to z. We will denote it by

$$\left(J \cdot j, z \cdot z^j\right) = (J \cup \{j\}, z')$$

where $z'^l = z^l$ for $l \in J$ and $z'^j = z^j$.

Clearly, a set $J \subseteq M$ defines $2^{|J|}$ such bins (many of which may be empty). A new point x_p corresponds to one such bin. The probabilistic prediction for y_p corresponding to x_p is the average frequency of 1's in it. If a bin is empty, this prediction is 0.5. Formally, the prediction is given by

$$\overline{y}^{(J,z)} = \frac{\sum_{i \in b(J,z)} y_i}{|b(J,z)|} \tag{3}$$

if $|b(J,z)| > 0$ and $\overline{y}^{(J,z)} = 0.5$ otherwise.

For the sake of calculating the empirical similarity, for each $i \leq n$ we consider the bin containing it, $b(J,z)$, and the value $\overline{y}_i^s$ is the average frequency of 1's in the bin once observation i has been removed from it. If $b(J,z) = \{i\}$, that is, the bin contains but one observation, taking one out leaves us with an empty database, resulting in a probabilistic prediction – and an error – of 0.5. Formally, the leave-one-out prediction for $i \in b(J,z)$ is

$$\overline{y}_i^{(J,z)} = \frac{\sum_{k \in b(J,z), k \neq i} y_k}{|b(J,z)| - 1} \tag{4}$$

if $|b(J,z)| > 1$ and $\overline{y}_i^{(J,z)} = 0.5$ otherwise.

Splitting the database into such bins is clearly an artifact of our definition of the similarity function, which renders the relation "having positive similarity" – equivalent to "having similarity 1" – an equivalence relation. This is a highly idealized model that is only used to convey some main features. We discuss extensions to more realistic, non-binary models, in Section **??**.

Choosing a subset of variables to be included in J is akin to selecting a subset of variables in a regression problem. There are two types of considerations for which smaller sets of predictors might be preferred to larger ones in both types of problems. The first, statistical considerations are normative in nature, and have to do with avoiding overfitting. The second are psychological, and have a descriptive flavor: people may not be able to recall and process too many variables. As a normative theory, the preference for simple theories is famously attributed to William of Ockam (though he was not explicitly referring to out-of-sample prediction errors), and runs throughout the statistical literature of the 20th century (see Akaike, 1974).[5] As a descriptive theory, the preference for simplicity appears in Wittgenstein's Tractatus (1923) at the latest. Moreover, one may argue that such preference is evolutionarily selected partly due to the statistical normative considerations.

Be that as it may, it seems likely that people would prefer a smaller set of predictors, given a fixed level of goodness of fit, and that they would even be willing to trade off the two.[6] We will capture this preference using the simplest model that conveys our point. Let us assume that the agent selects a similarity function that minimizes an *adjusted mean squared error*. Formally, the agent is assumed to select a set of indices J that minimizes

$$AMSE(J,c) \equiv MSE(J) + c|J|$$

for some $c \geq 0$. We will typically think of c as small, so that goodness of fit

[5] Note, however, that no matter how many variables one includes, a perfect fit isn't guaranteed in our case. The variables are only used to determine similarity weights, while prediction remains a function of observed values of y, and not of the x's directly.

[6] As we will shortly, for case-based prediction the minimization of the SSE may favor smaller sets of predictors even without the introduction of preference for simplicity.

would outweigh simplicity as theory selection criteria, but as positive, so that complexity isn't ignored. Given a cost c, we will refer to a similarity function $s = s_{w_J}$ for $J \in \arg\min AMSE(J, c)$ as *an empirical similarity function.*

3 General Properties

In this section we discuss a few general properties of the empirical similarity. First, we point out that even if $c = 0$, the empirical similarity may not favor more variables to less. In other words, a set of variables may have a higher MSE than a subset thereof. We then present conditions under which monotonicity of the MSE with respect to set inclusion is guaranteed. Second, we point out that the empirical similarity may not be unique. Further, we show that it is hard to compute. We conclude discussing how our results imply that different people can use different similarity functions even if they are exposed to the same database and use the same belief-formation process.

3.1 Monotonicity

We start by showing that using a relatively small set of variables for prediction might be desirable even with $c = 0$, because the goodness-of-fit (for a given database) can *decrease* when adding one more predictor.[7] The reason is a version of the problem known as "the curse of dimensionality": more variables that are included in the determination of similarity would make a given database more "sparse". The following example illustrates.

Example 1 Let $n = 4$ and $m = 1$ with the following database:

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

[7] Notice that this cannot happen with other statistical techniques such as linear regression.

It can be seen that the MSE's of the subsets of variables are given by

J	$MSE\,(J)$
$\varnothing$	$4/9$
$\{1\}$	1

The specific form of the curse of dimensionality that affects the leave-one-out criterion is due to the fact that this criterion compares each observation (y) to the average of the *other* observations. A bin that contains $a > 0$ cases with $y_i = 1$ and $b > 0$ cases with $y_i = 0$ has an average y of $\frac{a}{a+b}$. But when an observation $y_i = 1$ is taken out, it is compared to the average of the remaining ones, $\frac{a-1}{a+b-1} < \frac{a}{a+b}$, and vice versa for $y_i = 0$ (which is compared to $\frac{a}{a+b-1} > \frac{a}{a+b}$). In both cases, the squared error in the leave-one-out computation is larger than it would be in the computation of the sample's variance (comparing y_i to $\frac{a}{a+b}$). Most relevantly, this squared error decreases in the size of the bin because the larger the bin, the smaller the impact of taking out a single observation on the average of the remaining ones.

The above suggests that in appropriately-defined "large" databases the curse of dimensionality would be less severe and adding variables to the set of predictors would be easier than in smaller databases. To make this comparison meaningful, and control for other differences between the databases, we can compare a given database with "replications" thereof, where the counters a and b above are replaced by ta and tb for some $t > 1$. Formally, we will use the following definition.

Definition 1 *Given two databases* $B = \{(x_i, y_i)\}_{i \leq n}$ *and* $B' = \{(x'_k, y'_k)\}_{k \leq tn}$ *(for* $t \geq 1$*), we say that* B' *is a* t-replica of B *if, for every* $k \leq tn$, $(x'_k, y'_k) = (x_i, y_i)$ *where* $i = k (\operatorname{mod} n)$.

We can now a database B' which is a t-replica of the database in Example 1. It can readily be verified that

$$MSE\,(\varnothing) = \left(\frac{2t}{4t-1}\right)^2 < \left(\frac{t}{2t-1}\right)^2 = MSE\,(\{1\})\,.$$

Indeed, the dramatic difference of the MSE's in Example 1 ($[MSE(\{1\}) - MSE(\varnothing)]$) is smaller for larger t's, and converges to 0 as $t \to \infty$. However, the non-monotonicity of the MSE with respect to set inclusion (that is, the fact that a set can have a higher MSE than a subset thereof) can still occur in databases that are large, both in terms of the overall number of cases in them as in terms of the number of observations in each bin.

This suggests that there is something special about Example 1 beyond the size of the database. Indeed, the variable in question, x^1, is completely uninformative: the distribution of y is precisely the same in each bin (i.e., for $x^1 = 0$ and for $x^1 = 1$), and thus there is little wonder that splitting the database into these two bins can only result in larger errors due to the bin sizes, with no added explanatory power to offset it. Formally, we define informativeness of a variable (for the prediction of y in a database B) relative to a set of other variables as a binary property:

Definition 2 *A variable $j \in M$ is* informative *relative to a subset $J \subseteq M \backslash \{j\}$ in database $B = \{(x_i, y_i)\}_{i \leq n}$ if there exists $z \in \{0,1\}^J$ such that $|b(J, z \cdot 0)|, |b(J, z \cdot 1)| > 0$ and*

$$\overline{y}^{(J \cdot j, z \cdot 0)} \neq \overline{y}^{(J \cdot j, z \cdot 1)}$$

In other words, a variable x^j is informative for a subset of the variables, J, if, for at least one assignment of values to these variables, the relative frequency of $y = 1$ in the bin defined by these values and $x^j = 1$ and the relative frequency defined by the same values and $x^j = 0$ are different.

One reason that a variable j may be uninformative relative to a set of other variables is that it can be completely determined by them. Consider the following two databases.

Example 2 Let $n = 2$ and $m = 2$ with the following databases:

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

In both examples 2 is uninformative relative to $\{1\}$ and vice versa, as the contain exactly the same information.

Formally,

Definition 3 *A variable* $j \in M$ is a function of $J \subseteq M \backslash \{j\}$ in database $B = \{(x_i, y_i)\}_{i \leq n}$ *if there is a function* $f : \{0,1\}^J \to \{0,1\}$ *such that, for all* $i \leq n$, $x_i^j = f\left(\left(x_i^k\right)_{k \in J}\right)$.

If j is a function of J, the bins defined by J and by $J \cup \{j\}$ are identical, and clearly j cannot be informative relative to J. However, as we saw above, a variable j may fail to be informative relative to J also if it isn't a function of J. To determine whether j is a function of J we need not consult the y values. Informativeness, by contrast, is conceptually akin to correlation of the variable x^j with y given the variables in J.

We can finally state conditions under which more variables are guaranteed to result in a lower MSE. Intuitively, we want to start by adding a variable that is informative (relative to those already in use), and to make sure that the database isn't split into too small bins. Formally,

Theorem 1 *Assume that* j *is informative relative to* $J \subseteq M \backslash \{j\}$ *in the database* $B = \{(x_i, y_i)\}_{i \leq n}$. *Then there exists a* $T \geq 1$ *such that, for all* $t \geq T$, *for a* t*-replica of* B, $MSE\left(J \cup \{j\}\right) < MSE\left(J\right)$. *Conversely, if* j *is not informative relative to* J, *then for any* t*-replica of* B, $MSE\left(J \cup \{j\}\right) \geq MSE\left(J\right)$, *with a strict inequality unless* j *is a function of* J.

We note in passing that informativeness of a variable does not satisfy monotonicity with respect to set inclusion:

Observation 1 *Let there be given a database $B = \{(x_i, y_i)\}_{i \leq n}$, a variable $j \in M$, and two subsets $J \subseteq J' \subseteq M \backslash \{j\}$. It is possible that j is informative for J, but not for J' as well as vice versa.*

Theorem 1 indicates that, if the database is large enough in the sense that its minimal (nonempty) bins are large, and *if we were to ignore the cost of additional variables*, informative variables should be expected to be included in the set of variables that minimizes the MSE. The observation above warns us that variables that are informative on their own (that is, relative to the empty set) may not be informative when taken in conjunction with others, as well as vice versa. Yet, especially if we think of the binary model as a simplification of more realistic set-ups, we could expect variables to be informative and remain so also in the presence of others.

This discussion raises the question, should we expect databases to be "large" in this sense? With m variables, one can have up to 2^m bins. If n is large relative to 2^m, one can expect many of the bins to be large. This isn't necessarily the case, as many observations may belong to some bins, leaving very few observations for the other bins. But it would still be the case that most observations belong to large bins, and therefore adding variables might well decrease the MSE. By contrast, if n is small relative to 2^m, one can expect many bins to be small, and, importantly, many observations to belong to small bins, where the curse of dimensionality can have a bite. In other words, the order of quantifiers matters. If we fix m and increase n, we could expect additional variables to reduce the MSE, but if we fix n and increase m, the opposite may be the case.

Naturally, there are problems in which it makes sense to assume that n is large relative to 2^m and problems where the converse is more realistic. For example, when tossing a coin over and over again one may not have many variables to measure, and additional observations can be obtained relatively easily. More generally, when experimentation is feasible and not too costly, one can add more datapoints whenever a new variable comes to mind. This, indeed, is expected of scientific studies, and in these cases one can expect all informative variables to be used. By contrast, if the observations involve past presidential elections, financial crises, wars, or periods of economic growth, one cannot add cases at will. Moreover, for the n cases given by history one can often come up with additional variables that may be relevant, making 2^m large relative to n, and resulting in a selection of a subset of the informative variables.

3.2 Uniqueness

We have seen in section 3.1 that monotonicity of the MSE is not generally guaranteed. We now explore how this may result in non-uniqueness of the empirical similarity function, whether one considers the simplifying assumption of $c(J) = 0$ or allows c to be positive to avoid overfitting and take into account the preference for simplicity. Consider the following example:

Example 3 Let $n = 12t$ (for $t \geq 1$) and $m = 2$, and let the observations be t replications of the following

i	x_i^1	x_i^2	y_i
1	1	0	0
2	1	0	1
3	0	1	0
4	0	1	1
5-8	0	0	0
9-12	1	1	1

For $t = 1$, the MSE's of the subsets of variables are given by

J	$MSE(J)$
$\varnothing$	0.297
$\{1\}$	0.2
$\{2\}$	0.2
$\{1,2\}$	0.333

and thus the minimizers of the MSE are (only) $\{1\}$ and $\{2\}$. For large enough t, the estimated probability in each bin depends on the observation that is being left out only to a negligible degree. One may verify that, as $t \to \infty$, the MSE's of the subsets of variables are given by

J	$MSE\,(J)$
$\varnothing$	0.250
$\{1\}$	0.138
$\{2\}$	0.138
$\{1,2\}$	0.083

If $c > 0.6$, the minimizers of $AMSE$ are $\{1\}$ and $\{2\}$ but not $\{1,2\}$.

As in the case of maximizing the adjusted R^2 in linear regression (or some other measure that trades off goodness-of-fit vs. the number of variables), we find that the minimizer of the $AMSE$ need not be unique, even for a large database. In the case of a small database, even minimizing the (unadjusted) MSE can result in non-uniqueness (as opposed to the generic case of maximizing the unadjusted R^2).[8]

Non-uniqueness is particularly likely to occur if we are dealing with a small set of observations, for which many possible variables might be relevant. Specifically, consider the following class of data generating processes. Let n be fixed and let m grow to infinity. Assume that for each new variable j, and for every $i \leq n$,

$$P\left(x_i^j = 1 \,\middle|\, x_k^l,\, l < j \text{ or } (l = j, k < i)\right) \in (\varepsilon, 1 - \varepsilon)$$

for a fixed $\varepsilon \in (0, 0.5)$. That is, we consider a rather general joint distribution of the variables $x^j = \left(x_i^j\right)_{i \leq n}$, with the only constraint that the probability of the next observed value, x_i^j, being 1 or 0, conditional on all past observed values, is uniformly bounded away from 0, where "past" is read to mean "an observation of a lower-index variable or an earlier observation of the same variable". (The uniformity is needed because $m \to \infty$. Note that for any x_i^j

[8] Non-uniqueness may also result from the cost c that appears in the $AMSE$. People may differ in their tolerance of inaccuracy and of complexity and may choose different ways to trade them off.

there are only finitely many possible value combinations of the variables x_k^l that precede x_i^j in the lexicographic ordering dictated by columns and rows.) For such a process we can state:

Proposition 1 For every $n \geq 4$ and every $\varepsilon \in (0, 0.5)$, if there are at least two cases with $y_i = 1$ and at least two with $y_i = 0$, then, as $m \to \infty$, the probability that there exist J,J' with $J \cap J' = \varnothing$ and $MSE\left(J\right) = MSE\left(J'\right) = 0$ tends to 1.

The data generating process discussed can be viewed as a model of a process in which people come up with additional possible predictors for a given phenomenon, while the number of past cases is limited. Examples such as presidential elections and revolutions have a number of relevant cases that is more or less fixed, but these cases can be viewed from new angles, by introducing new variables that might be pertinent. The proposition suggests that, when more and more variables are considered, we should not be surprised if completely different (that is, disjoint) sets of variables obtain perfect fit.

3.3 Complexity

Examples in which different sets of variables obtain precisely the same $AMSE$ might be knife-edge, especially if we take the binary model as a metaphor for continuous ones. However, when the number of variables grows, so does the complexity of finding the optimal set of variables, even if it is unique. We first establish this fact formally and then discuss its implications.

Define the following problem:

Problem 1 EMPIRICAL-SIMILARITY: Given integers $m, n \geq 1$, a database $B = \{(x_i, y_i)\}_{i \leq n}$, and (rational) numbers $c, R \geq 0$, is there a set $J \subseteq M \equiv \{1, ..., m\}$ such that $AMSE(J, c) \leq R$?

Thus, EMPIRICAL-SIMILARITY is the yes/no version of the optimization problem, “Find the empirical similarity for database B and constant c”.

We can now state

Theorem 2 *EMPIRICAL-SIMILARITY is NPC.*

It follows that, when many possible variables exist, we should not assume that people can find an (or the) empirical similarity. That is, it isn't only the case that there are 2^m different subsets of variables, and therefore as many possible similarity functions to consider. There is no known algorithm that can find the optimal similarity in polynomial time, and it seems safe to conjecture that none would be found in the near future. [9]

We emphasize that the practical import of this complexity result depends crucially on the number of variables, m.[10] For example, if $m = 2$ and there are only 4 subsets of variables to consider, it makes sense to assume that people find the "best" one. Moreover, if n is large, the best one may well be all the informative variables. If we consider the case of scientific experimentation again, where additional experiments are feasible and not too costly, it stands to reason that non-uniqueness will not be an issue, and that there would be a high degree of convergence among rational agents who have access to the same database. By contrast, when there are more variables, even if the database is large, disagreement is possible. This might be the case in particular where data cannot be obtained at will. In many problems from the medical or educational domain there are many possible variables of interest (thus, exponentially many subsets of variables to consider), and, while there are many data, many experiments are ruled out by practical and ethical considerations. Finally, when considering historical events such as wars and revolutions, n is fixed and one can easily come up with a large number of variables that may potentially be informative. In these cases it seems rather

[9] This result is the equivalent of the main result in Aragones et al. (2005) for regression analysis. Thus, both in rule-based models and in case-based models of reasoning, it is a hard problem to find a small set of predictors that explain the data well.

[10] Indirectly, it also depends on n. If n is bounded, there could be only a bounded number (2^n) of different variable values, and additional ones need not be considered.

likely that individuals would disagree on the way similarity should be judged and probability assessed.

3.4 Different Beliefs

The results in sections 3.2 and 3.3 suggest two reasons why people may entertain different beliefs about future observations, even if they have access to the same database of cases and they employ the same method of belief generation. First, the empirical similarity might not be unique. For example, consider the database

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

where both $J = \{1\}$ and $J' = \{2\}$ obtain a minimal $AMSE(K, c)$ of c (and $AMSE(\{1, 2\}, c) = 2c$). If the next observation has $x = (1, 0)$ agents using J would predict $y = 1$ and agents using J' – $y = 0$.

As Proposition 1 demonstrates, such different of opinion is particularly common in cases such as presidential elections and revolutions where the a number of relevant cases is more or less fixed, but people can come up with many new potentially relevant predictors.

Second, finding the empirical similarity function is computationally complex. Therefore, when confronted with a database, people may resort to various heuristics in search of good similarity functions, but none of these heuristics is guaranteed to settle on the optimal function even if such exists. In particular, people may find "local optima", say, functions that cannot be improved upon by adding or dropping a single variable, but they may not be aware of other sets that require dropping m_1 variables and adding m_2 other variables instead. As there may be many local optima even generically, one should not marvel at the fact that different people may use different sets of variables to judge similarities and to make predictions.

4 Application: Investing in Precedents to Establish a Reputation

Consider a data-generating process such that beliefs about the outcome y_i are self-fulfilling because y_i represents the equilibrium of a coordination game. For example, y_i could represent the success or failure of a revolutionary attempt: only if sufficiently many people believe the attempt can be successful, will they take part in it and thereby bring about its success. Now suppose an agent, new to the scene, has an interest in affecting future beliefs, in order to determine future outcomes. Suppose that the new agent, whose identity can be modelled by the value 1 for a variable x_i^j, prefers the outcome to be $y_i = 1$. In the example, suppose that a new government wants to avoid future revolutions by convincing people that revolutions are unlikely to succeed under its "iron fist" rule. If beliefs are formed using the type of second-order induction discussed in this paper, the agent can manipulate them by investing resources to affect the outcome y_i for some periods, establishing precedents that change the empirical similarity. In the example, the government can invest resources in one or more conflicts, to make sure early revolutionary attempts fail. In this section we investigate the empirical similarity given precedents, and focus on the question, how many realizations of the outcome $y_i = 1$ are needed to affect future beliefs, thus creating a valuable "reputation" for the new agent.

More concretely, assume that there is a database of similar past cases, none of which includes the new agent, and for each of them it is known whether a certain outcome resulted ($y = 1$ or $y = 0$). We assume that all the relevant past cases, and all future cases to be discussed share the values of all variables in a set J, and they only differ in the value of a new variable, $j \notin J$, designating the agent's proper name. In all past cases $x_i^j = 0$ and we are interested in the empirical similarity that will be computed for various continuations of the database by additions of cases in which $x_i^j = 1$.

We consider here the simple case in which the database contains the same numbers of 0's and 1's, so that the probability that $y = 1$ next is estimated at 0.5. Thus, assume that in the past cases (where $x_i^j = 0$) we observed $y_i = 1$ N times, and $y_i = 0$ – N times. We consider the new agent who faces case $i = 2N + 1$, but takes into account future cases as well. We would therefore like to consider the empirical similarity that would be computed from the database in the future, where the first $2N$ cases are augmented by additional ones, in which $x_i^j = 1$. Let us therefore consider a database that contains $2N + k + l$ cases, of which $k \geq 0$ have $x_i^j = 1$ and $y_i = 1$, and $l \geq 0$ – $x_i^j = 1$ and $y_i = 0$.

We are interested in the question, how many times would the agent need to bring about the outcome $y_i = 1$ so that its proper name be part of the empirical similarity, and suggest the prediction that $y_{n+1} = 1$ is likely to be observed (in the next case). Note that, if $MSE\left(J \cup \{j\}\right) < MSE\left(J\right)$ and $k > l$, the variable j is included in the set of predictors, and it would make the observation $y_{n+1} = 1$ more likely than $y_{n+1} = 0$. Thus we fix N, and l, and ask what is the minimal $k = k\left(N, l\right) > l$ for which $MSE\left(J \cup \{j\}\right) < MSE\left(J\right)$.[11] That is, when will it be the case that people who have access to the database are convinced that this particular agent is in a class of it own.

We can now state

Proposition 2 Let there be given $N > 2$ and $l \geq 0$. Then:

(i) For every $l \geq 0$ there exists k_0 such that for all $k \geq k_0$, $MSE\left(J \cup \{j\}\right) < MSE\left(J\right)$; in particular, $k\left(N, l\right)$ is finite (as $k\left(N, l\right) \leq k_0$);

(ii) $k\left(N, 0\right) = 2$;

(iii) $k\left(N, 1\right) = 5$.

This proposition shows the asymmetry between establishing and losing

[11] Clearly, in this case minimizing the $AMSE$ for low enough cost c will also favor the inclusion of j in the similarity function.

reputation. Starting with a blank slate, if the agent can set $y_{n+1} = y_{n+2} = 1$, then the MSE is minimized with the inclusion of j in the similarity function. This means that the agent managed to establish reputation, and that other agents think of her as a class apart. Naturally, when more such cases ($y_i = 1$) are accumulated, this reputation is further established. However, if there is but one case of $y_i = 0$ (say, $i = n+1$), even two additional cases of $y_i = 1$ will still not suffice to re-established reputation. In particular, reputation that was obtained in two cases can be destroyed by a single case, and it would take at least five additional cases to re-establish it.

4.1 Example: The Collapse of the USSR

The Soviet bloc started collapsing with Poland, which was the first country in the Warsaw Pact to break free from the rule of the USSR. Once this was allowed by the USSR, other countries soon followed. One by one practically all the USSR satellites in Eastern Europe underwent democratic revolutions, culminating in the fall of the Berlin Wall in 1989.

Revolutions are often seen as a change of equilibrium. Further, it has been argued that similarity-weighted frequencies of past cases can be applied to the prediction of a success of a possible revolution, and therefore also to the prediction of revolution attempts (see Steiner and Stewart, 2008, Argenziano and Gilboa, 2012). It appears obvious that the case of Poland was an important precedent, which generated a "domino effect". According to our model, its importance didn't lie only in changing the relative frequencies, but also via second-order induction, dropping the attribute "being a part of the Soviet Bloc" from the empirical similarity function.

Similarly, when the Baltics were allowed to secede from the USSR in 1991, the USSR disintegrated. This can be viewed as another change in the similarity function: the attribute "being a part of the USSR", which separated the Baltics from Poland, was no longer deemed relevant. Soon after, Chechnya attempted to claim independence from Russia. A success would have proven that even the variable “being a part of Russia” was no longer relevant. This, apparently, was not something Russia could afford. Thus, one could view the battle over Chechnya as a conflict over future empirical similarity.[12]

4.2 Example: Currency Change

In an attempt to restrain inflation, central banks sometimes resort to changing the currency. France changed the Franc to New Franc (worth 100 “old” francs) in 1960, and Israel switched from a Lira to a Shekel (worth 10 Liras) in 1980 and then to a New Shekel (worth 1,000 Shekels) in 1985.

A change of currency has an effect at the perceptual level of the similarity function. Different denominations might suggest that the present isn’t similar to the past, and that the rate of inflation might change. However, if people engage in second-order induction, they would observe new cases and would learn from them whether the perceptual change is of import. For example, the change of currency in Israel in 1980 was not accompanied by policy changes, and inflation spiraled into hyper-inflation. By contrast, the change in 1985 was accompanied by budget cuts, and inflation was curbed. The contrast between these two examples suggests that economic agents are sufficiently rational to engage in learning the empirical similarity.

5 Generalizations

Our model can be extended in several ways, allowing the predictors $(x^1, ..., x^m)$ and/or the predicted variable y to be non-binary, and, importantly, the weights of attributes in the similarity function to be arbitrary non-negative numbers. Some of these generalizations would involve new insights. For example, by changing the empirical similarity, cases can affect the predictions

[12] If different variables are dropped (or added) to the similarity function consecutively one may identify such a pattern and think of it as “third-order induction”. Our formal model does not extend thus far.

for other cases without necessarily being similar to them.

Consider the motivating example again. We argued that the precedent of President Obama reduced the importance of the variable “race” in similarity judgments. This may make other African Americans more likely to win an election for two reasons: first, they are similar to the precedent; second, the attribute on which they differ from the vast majority of past cases is less important. With variables that can take more than two values, one can have the latter effect without the former. Suppose that, in an upcoming election, an American-born man of Chinese origin considers running for office. If, indeed, the empirical similarity function does not put much weight on the variable “race”, such a candidate would be more likely to win an election, after the case of Obama, without necessarily being similar to the latter.[13]

One can also extend the model to deal with continuous variables, allowing the predictors $(x^1, ..., x^m)$ to assume values (jointly) in a set $X \subseteq \mathbb{R}^m$ while the predicted variable, y, – in a set $Y \subseteq \mathbb{R}$. It is natural to use the same formulae of similarity-weighted average used for the binary case, i.e.,

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)} \tag{5}$$

this time interpreted as the predicted value of y (rather than the estimation of the probability that it be 1). This formula was axiomatized in Gilboa, Lieberman, and Schmeidler (2006).[14]

For many purposes it makes sense to consider more general similarity

[13] This prediction of our model could be tested empirically. Admittedly, should it prove correct, one could still argue that the similarity function has a variable “Non-Caucasian” (rather than “race”), so that a Chinese-born and an African-American are similar to each other in this dimension. We find the change of the similarity function to be a more intuitive explanation.

[14] If Y is discrete, we may also define the predicted value of y_p by

$$\hat{y}_p^s \in \arg\max_y \sum_{i \leq n} s(x_i, x_p) \mathbf{1}_{\{y = y_i\}} \tag{6}$$

which is equivalent to kernel classification and has been axiomatized in Gilboa and Schmeidler (2003).

functions, that would allow for values in the entire interval $[0,1]$ and would not divide the database into neatly separated bins. In particular, Billot, Gilboa, and Schmeidler (2008) characterize similarity functions of the form

$$s\left(x,x'\right)=e^{-n\left(x,x'\right)}$$

where n is a norm on $\mathbb{R}^m$. Gilboa, Lieberman, and Schmeidler (2006) and Gayer, Gilboa, Lieberman (2007) also study the case of a weighted Euclidean distance, where

$$s\left(x,x'\right)=\exp\left(-\sum_{j=1}^{m}w^j\left(x^j-x'^j\right)^2\right) \tag{7}$$

with $w_j \geq 0$.[15]

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 $s\left(x,x'\right)=0$ whenever $x^j\neq x'^j$, but if $x^j=x'^j$, the j-th summand in (7) will be taken to be zero. In other words, we allow for the value $w^j=\infty$ with the convention that $\infty\cdot 0=0$. This would make the binary model a special case of the current one. (Setting $w^j=\infty$ in (7) where $w^j=1$ in (2).) For the computational model, the value ∞ will be considered an extended rational number, denoted by a special character (say "∞"). The computation of $s\left(x,x'\right)$ 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 $s\left(x,x'\right)=0$. Otherwise, the computation proceeds with (7) where the summation is taken over all j's such that $w^j<\infty$.

The definition of the empirical similarity extends to this case verbatim: the SSE and the MSE are defined in the same way, and one can consider similarity functions given by (7) for some non-negative $\left(w^j\right)_{j\leq m}$. Rather than thinking of $SSE\left(s\right)$ as a function of a set of predictors, $J\subseteq M$, denoted

[15] If one further assumes that there is a similarity-based data generating process driven by a function as the above, one may test hypotheses about the values of the weights w_j.

$SSE\,(J)$ as above, one would consider it as a function of a vector of weights, $w = \left(w^j\right)_{j\leq m}$, denoted $SSE\,(w)$.

It is natural to define the $AMSE$ by

$$AMSE(w,c) \equiv MSE(w) + c|J\,(w)\,|$$

where

$$J\,(w) = \left\{\, j \leq m \;\middle|\; w^j > 0 \,\right\}.$$

That is, a positive weight on a variable incurs a fixed cost. This cost can be thought of as the cost of obtaining the data about the variable in question, as well as the cognitive cost associated with retaining this data in memory and using it in calculations.

We argue that the main message of our results in the binary case carry over to these models as well. Clearly, this would depend on the assumptions one imposes on the data generating process. For example, if the variables are drawn from a continuous distribution, one would expect that, generically, the empirical similarity function would be unique. Yet, it appears that the behavior of the SSE as a function of s is hardly conducive to optimization. There is no reason to assume that the SSE is a convex or a quasi-convex function, and casual observation suggests that the SSE may have multiple local minimizers. Importantly, the our complexity result extends to this case. Formally,

Problem 2 CONTINUOUS-EMPIRICAL-SIMILARITY: 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 *CONTINUOUS-EMPIRICAL-SIMILARITY is NPC.*

As will be clear from the proof of this result, the key assumption that drives the combinatorial complexity is not that x, y or even w are binary. Rather, it 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 $w^j = 0$.[16][17]

To conclude, it appears that the qualitative conclusion, namely that people may have the same database of cases yet come up with different "empirical similarity" functions to explain it, would hold also in a continuous model.

6 Discussion

6.1 Compatibility with Bayesianism

There are (at least) three ways in which one can apply the Bayesian approach to the problems we consider. Learning the similarity function is compatible with two of them, but not with the third:

– One may adopt a "small world" approach, and try to develop a prior probability for a particular case, such as a single revolution attempt. In the examples discussed above this "prior" would be summarized by a single probability number, and there wouldn't be any opportunity to perform Bayesian updating. One may develop slightly more elaborate models, in which each case would involve a few stages (say, demonstrations, reaction by the regime, siege of parliament...) and use past cases to define a prior on the multi-stage space, which can be updated after some stages have been observed. Our approach is compatible with this version of Bayesianism, where the similarity-based relative frequencies using the empirical similarity is a

[16] To see that this complexity result does not hinge on specific values of the variables x_i^j and each y_i, one may prove an analogous result for a problem in which positive-length *ranges* of values are given for these variables, where the question is whether a certain $AMSE$ can be obtained for some values in these ranges.

[17] See also Eilat (2007), who 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 Aragones et al., 2005).

method of generating a prior belief over the stages of the attempted revolution.

– Alternatively, one can adopt a "large world" or "grand state space" approach, in which a state of the world resolves any uncertainty from the beginning of time. Savage (1954) suggests to think of a single decision problem in one's life, as if one were choosing a single act (strategy) upon one's birth. Thus, the newborn baby would need to have a prior over all she may encounter in her lifetime. For many applications one may need to consider historical cases, and thus the prior should be the hypothetical one the decision maker would have had, had she been born years back. The assumption that newborn entertain a prior probability over the entire paths their lives would take is a bit fanciful. Further, the assumption that they would have such a prior even before they could make any decisions conflicts with the presumably-behavioral foundations of subjective probability. Yet, this approach is compatible with the process we describe: in the language of such a model, ours can be described as agents having a high prior probability that the data generating process would follow the empirical similarity function. In the context of a game (such as a revolution), this would imply that they expect other players' beliefs to follow a similar process.

– There are ways of implementing the Bayesian approach that are in between the small world and the large world interpretation, and these are unlikely to be compatible with our model. For example, assume that an agent believes that the successes of revolutions generates a (conditionally) i.i.d. sequence of Bernoulli random variables, with an unknown parameter p. As a Bayesian statistician, she has a prior probability over p, and she observes past realizations in order to infer what p is likely to be. This Bayesian updating of the prior over p to a posterior has no reason to resemble our process of learning the similarity function.

Our main interest is in problems that do not readily yield to statistical analysis. The assumption that revolutions, wars, or financial crises are i.i.d.

(conditional on a set of parameters) seems hardly tenable, especially because these phenomena seem to be strongly causally related. One revolution may lead to another, via a "domino effect", whereas a financial crisis may obviate another, say, due to central banks' reactions. Thus, we do not find the third Bayesian approach of great appeal for the problems we have in mind. By contrast, one can certainly view the process we describe as a way in which agents generate prior beliefs for a "small world". We would view such a belief generation process as reasonably rational, and, in the case of a game, also as an equilibrium in the belief-generation meta-game. That is, if players who are engaged in, say, a coordination game first choose a method for generating beliefs, it will be an equilibrium for them to generate beliefs in the same way. In this game, the empirical similarity function may be a focal point, suggesting particular beliefs as the equilibrium choice.

6.2 Compatibility with Common Knowledge of Equilibrium Selection

The discussion of reputation assumed that a large population of players is involved in a coordination game, and that the equilibrium selection is done by the prediction of play using the empirical similarity. We have argued above that computing the empirical similarity is compatible with a Bayesian approach applied to the grand state space; but is it also compatible with common knowledge of the rationality of others, and of the model itself?

The answer depends on the interpretation of the computation of the empirical similarity. If we assume that each player believes that the real process is governed by a fixed similarity function, and tries to learn it by minimizing the $AMSE$, there seems to be a conflict between the modeler's world view and those of the players: according to this interpretation, the modeler is the only one who knows that all the players are computing the empirical similarity function, while each of them assumes that she is the only one to do so, and that the others are not as sophisticated.

But one may also adopt an interpretation that would make the model, and rationality of each agent therein, common knowledge: rather than believing that the process is governed by a fixed similarity function, one can think of the empirical similarity calculation merely as a focal point on which the players converge. Indeed, in a coordination game any algorithm for generating beliefs can serve as a coordinating device. We can think of an implicit pre-play game, in which players choose their beliefs about the coordination game. This pre-play game would also be a coordination game, and any method for belief generation would suggest an equilibrium.

With this interpretation in mind, we suggest that the players use the empirical similarity prediction as a focal point because it is a reasonable process in non-strategic setups. By analogy, consider a game in which players observe k rolls of a die, and then have to select points in $\Delta \equiv \Delta\left(\{1, 2, ..., 6\}\right)$. A player who picks $p_i \in \Delta$ gets a payoff $-\left\|p_i - \bar{p}\right\|$ where $\bar{p}$ is the average of the p_i's. Clearly, the selection $p_i = \hat{p}$ (for all i) is an equilibrium for any $\hat{p} \in \Delta$. Yet, the empirical frequency of the rolls of the die seems to stand out as a focal point. One reason might be that the empirical frequency would be a good guess in a prediction problem where the process i.i.d. Relatedly, if some of the players mistakenly ignore the strategic aspect of the game and focus on prediction the next observation, then (with a quadratic loss function) they would select the empirical frequency. If other players are trying to minimize the loss function with the realization that some of the players are non-strategic, they might also select the empirical frequency.

Along similar lines, in our case there may be some players who are non-strategic and do indeed believe that the process follows an unknown similarity function. Attempting to estimate this function, they would use the empirical similarity to generate their prediction over the game's equilibrium. Other players might engage in Level-1 reasoning, and optimally react to the existence of Level-0 players by predicting the equilibrium chosen by the empirical similarity. As this is a coordination game, best response implies behavior according to the Level-0 beliefs. Similarly, higher levels of reasoning would also follow the same equilibrium prediction. In other words, the Level-0 prediction, which is statistically but not strategically sophisticated, isn't only a reasonable focal point in the belief-selection equilibrium; it is also the best prediction of the strategic choices of players who engage in Level-k reasoning for any $k \geq 0$.

6.3 Cases and Rules

As mentioned in the context of our examples above, one can assume that people use rule-based, rather than case-based reasoning, and couch the discussion in the language of rules. In particular, rules are naturally learnt from the data by a process of abduction (or case-to-rule induction), and a precedent can be viewed as a counter-example to a rule.

While the two modes of reasoning can sometimes be used to explain similar phenomena, they are in general quite different. First, sets of rules may be inconsistent, whereas this is not a concern for databases of cases. Second, association rules such as "if x_i belongs to a set..., then y_i is..." do not have a bite where their antecedent is false. Finally, association rules, which are natural for deterministic predictions, need to be augmented in order to generate probabilities.

We find case-based reasoning to be simpler for our purposes. Cases never contradict each other; their similarity-weighted relative frequency always defines a probability; and, importantly, they are a minimal generalization of simple relative frequencies that used to define objective probabilities. However, additional insights can be obtained from more general models that combine case-based and rule-based reasoning, with second-order induction processes that learn the similarity of cases as well as the applicability and accuracy of rules.

7 Appendix A: Proofs

Proof of Theorem 1:

Assume first that $j \in M$ is informative relative to $J \subseteq M \backslash \{j\}$ in $B = \{(x_i, y_i)\}_{i \leq n}$. Let $z \in \{0,1\}^J$ be such that $|b(J, z \cdot 0)|, |b(J, z \cdot 1)| > 0$ and

$$\overline{y}^{(J \cdot j, z \cdot 0)} \neq \overline{y}^{(J \cdot j, z \cdot 1)}$$

Assume that B' is a t-replica of B. The main point of the proof is that, for large enough t, the MSE of a given subset of variables, L, could be approximated by a corresponding expression in which $\overline{y}_i^{(L,z)}$ (computed for the bin in which i was omitted) is replaced by $\overline{y}^{(L,z)}$ (computed for the bin without omissions), and then to use standard analysis of variance calculation to show that the introduction of an informative variable can only reduce the sum of squared errors.

Formally, let $b_t(L, z')$ denote the L-z' bin in B' (so that $|b_t(L, z')| = t\,|b(L, z')|$). Recall that

$$MSE(L) = \frac{1}{n} \sum_{z' \in \{0,1\}^L} \sum_{i \in b_t(L,z')} \left(\overline{y}_i^{(L,z')} - y_i \right)^2$$

and define

$$MSE'(L) = \frac{1}{n} \sum_{z' \in \{0,1\}^L} \sum_{i \in b_t(L,z')} \left(\overline{y}^{(L,z')} - y_i \right)^2 .$$

We wish to show that these two expressions are close when t is large. Consider a particular bin $b_t(L, z')$. Assume that, for the original database B, the bin $b(L, z')$ is non-empty and has $a \geq 0$ observations with $y = 1$ and $b \geq 0$ with $y = 0$, such that $a + b = |b(L, z')| > 0$. In that case

$$\overline{y}_i^{(L,z')} = \begin{cases} \frac{a-1}{a+b-1} & y_i = 1 \\ \frac{a}{a+b-1} & y_i = 0 \end{cases}$$

while

$$\overline{y}^{(L,z')} = \frac{a}{a+b}.$$

If $a = 0$ or $b = 0$, then for any $t > 1$,

$$\sum_{i \in b_t(L,z')} \left(\overline{y}_i^{(L,z')} - y_i \right)^2 = \sum_{i \in b_t(L,z')} \left(\overline{y}^{(L,z')} - y_i \right)^2 = 0.$$

Consider, then, the case $a, b > 0$. Then

$$\left| \overline{y}_i^{(L,z')} - \overline{y}^{(L,z')} \right| = \begin{cases} \frac{bt}{(at+bt)(at+bt-1)} & y_i = 1 \\ \frac{at}{(at+bt)(at+bt-1)} & y_i = 0 \end{cases}$$

and thus $\overline{y}_i^{(L,z')} - \overline{y}^{(L,z')} = O\left(\frac{1}{t}\right)$.

Next, observe that

$$\begin{aligned} & \left(\overline{y}_i^{(L,z')} - y_i \right)^2 - \left(\overline{y}^{(L,z')} - y_i \right)^2 \\ = & \left(\overline{y}_i^{(L,z')} - \overline{y}^{(L,z')} \right) \left(\overline{y}_i^{(L,z')} + \overline{y}^{(L,z')} - 2y_i \right) \end{aligned}$$

which implies that

$$\sum_{i \in b_t(L,z')} \left(\overline{y}_i^{(L,z')} - y_i \right)^2 - \sum_{i \in b_t(L,z')} \left(\overline{y}^{(L,z')} - y_i \right)^2 = O(1)$$

and

$$MSE(L) - MSE'(L) = O\left(\frac{1}{t}\right). \tag{8}$$

Let us now consider the given set of variables J and $j \in M \backslash J$ that is informative relative to J. For any $z' \in \{0,1\}^J$ we have (the standard analysis of variance calculation):

$$\begin{aligned} & \sum_{i \in b(J,z')} \left(\overline{y}^{(J,z')} - y_i \right)^2 \\ = & \left[\sum_{i \in b(J,z'), x_i^j = 0} \left(\overline{y}^{(J \cdot j, z' \cdot 0)} - y_i \right)^2 \right] + b \left(\overline{y}^{(J \cdot j, z' \cdot 0)} - \overline{y}^{(J,z')} \right)^2 \\ & + \left[\sum_{i \in b(J,z'), x_i^j = 1} \left(\overline{y}^{(J \cdot j, z' \cdot 1)} - y_i \right)^2 \right] + a \left(\overline{y}^{(J \cdot j, z' \cdot 1)} - \overline{y}^{(J,z')} \right)^2 \end{aligned}$$

where $a = \left|\left\{i \in b(J,z') \,\middle|\, x_i^j = 1\right\}\right|$ and $b = \left|\left\{i \in b(J,z') \,\middle|\, x_i^j = 0\right\}\right|$. Hence

$$\begin{aligned}\sum_{i\in b(J,z')} \left(\overline{y}^{(J,z')} - y_i\right)^2 &= \sum_{i\in b(J,z')} \left(\overline{y}^{\left(J\cdot j,z'\cdot x_i^j\right)} - y_i\right)^2 \\ &\quad + a\left(\overline{y}^{(J\cdot j,z'\cdot 1)} - \overline{y}^{(J,z')}\right)^2 + b\left(\overline{y}^{(J\cdot j,z'\cdot 0)} - \overline{y}^{(J,z')}\right)^2.\end{aligned}$$

It follows that, for every z', $\sum_{i\in b(J,z')} \left(\overline{y}^{(J,z')} - y_i\right)^2 \geq \sum_{i\in b(J,z')} \left(\overline{y}^{\left(J\cdot j,z'\cdot x_i^j\right)} - y_i\right)^2$, and for z (for which $\overline{y}^{(J\cdot j,z\cdot 0)} \neq \overline{y}^{(J\cdot j,z\cdot 1)}$ is known),

$$\sum_{i\in b(J,z)} \left(\overline{y}^{(J,z)} - y_i\right)^2 > \sum_{i\in b(J,z)} \left(\overline{y}^{\left(J\cdot j,z\cdot x_i^j\right)} - y_i\right)^2 + (a+b)\,c$$

where

$$c = \frac{a}{a+b}\left(\overline{y}^{(J\cdot j,z\cdot 1)} - \overline{y}^{(J,z)}\right)^2 + \frac{b}{a+b}\left(\overline{y}^{(J\cdot j,z\cdot 0)} - \overline{y}^{(J,z)}\right)^2 > 0.$$

Further, $c > 0$ is independent of t (as the averages in each bin do not change by replication). It follows that

$$MSE'\left(J \cup \{j\}\right) \leq MSE'\left(J\right) - c'$$

where $c' = \frac{|b(J,z)|}{n} > 0$ is independent of t. This, combined with (8), means that $MSE\left(J \cup \{j\}\right) < MSE\left(J\right)$ holds for large enough t.

Conversely, if j is not informative relative to J, then it remains non-informative for any t-replica of B. If j is a function of J, then the J bins and the $J \cup \{j\}$-bins are identical, with the same predictions and the same error terms in each, so that $MSE\left(J \cup \{j\}\right) = MSE\left(J\right)$. Assume, then, that j is not informative relative to J (for B and for any replica thereof), but that j isn't a function of J. Thus, at least one J-bin of B, and of each replica thereof, B', is split into two $J \cup \{j\}$-bins, but the average values of y in any two such sub-bins are identical to each other. It is therefore still true that $MSE'\left(J \cup \{j\}\right) = MSE'\left(J\right)$ because the sum of squared errors has precisely the same error expressions in both sides. However, for every set of variables L and every L-bin in which there are both $y_i = 1$ and $y_i = 0$, the error terms for that bin in $MSE\left(L\right)$ are higher than those in $MSE'\left(L\right)$: the leave-one-out technique approximates $y_i = 1$ by $\overline{y}_i^{(L,z')} < \overline{y}^{(L,z')}$ and $y_i = 0$ by $\overline{y}_i^{(L,z')} > \overline{y}^{(L,z')}$. Further the difference $\left|\overline{y}_i^{(L,z')} - \overline{y}^{(L,z')}\right|$ decreases monotonically in the bin size. Therefore, if at least one J-bin is split into two $J \cup \{j\}$-bins, we obtain $MSE\left(J \cup \{j\}\right) > MSE\left(J\right)$. □

Proof of Observation 1:

Consider a database obtained by $t > 1$ replications of the following ($n = 4t$, $m = 3$):

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

Clearly, y is a function of (x^1, x^2). In fact, it is the exclusive-or function, that is $y = 1$ iff $x^1 = x^2$. Neither 1 nor 2 is informative relative to $\varnothing$, but each is informative relative to the other. (Thus, for $J \equiv \varnothing \subseteq J' \equiv \{2\}$, $j = 1$ is informative relative to J' but not relative to J.) However, 1 is not informative relative to $J'' = \{2, 3\}$ (while it is relative to its subset J').

To see that the latter can happen also when the variable in question isn't a function of the other ones, consider the following example. Consider $n = 15$,$m = 2$:

i	x_i^1	x_i^2	y_i
1	0	0	0
2	0	0	1
3-6	0	1	0
7-8	0	1	1
9-10	1	0	0
11-12	1	0	1
13-14	1	1	0
15	1	1	1

It can be verified that x^1 is informative relative to $\varnothing$ but not relative to $\{2\}$. □

Proof of Proposition 1:

As there are at least two observations with the value of $y_i = 0$ and at least two with $y_i = 1$, if there is a variable j such that $x_i^j = y_i$ (or $x_i^j = 1 - y_i$) for all $i \leq n$, the set $J = \{j\}$ obtains $MSE(J) = 0$ (and $AMSE(J) = c$). We will show that the proposition holds for J and J' that are singletons.

Let the variables be generated according to the process described with $0 < \varepsilon < 0.5$. Each x^j has a probability of equalling y that is at least ε^n. The probability it does *not* provide a perfect fit is bounded above by $(1 - \varepsilon^n) < 1$ – which is a common bound across all possible realizations of previously observed variables. The probability that none of m such consecutively drawn variables provides a perfect fit is bounded above by $(1 - \varepsilon^n)^m \to 0$ as $m \to \infty$. Similarly if we consider $m = 2k$ variables, and ask what is the probability that there is at least one among the first k and at least one among the second k such that each provides a perfect fit ($x_i^j = y_i$ for all i) is at least $[1 - (1 - \varepsilon^n)^m]^2 \to 1$ as $m \to \infty$. $\square$

Proof of Theorem 2:

Clearly, EMPIRICAL-SIMILARITY is in NP. Given a set of variable indices, $J \subseteq M \equiv \{1, ..., m\}$, computing its AMSE takes no more than $O(n^2 m)$ steps.

The proof is by reduction of the SET-COVER problem to EMPIRICAL-SIMILARITY. The former, which is known to be NPC (see Garey and Johnson, 1979), is defined as

Problem 3 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, $c = (mn^3)^{-1}/2$ and $R = kc$. This construction can obviously be done in polynomial time.

We claim that there is a cover of size k of P iff there is a similarity function defined by a subset $J \subseteq M \equiv \{1, ..., m\}$ such that $AMSE(J, c) \leq R$. Let us begin with the "only if" direction. Assume, then, that such a cover exists. Let J be the indices $1 \leq i_1 \leq i_2 \leq ... \leq i_k \leq r = m$ of the cover. For every $i \leq 2p$, there exists $j \in J$ such that $x_i^j = 1$ and thus i is not in the same bin as $2p+1, 2p+2$. It follows that for every i' such that $s_{w_J}(x_i, x_{i'}) = 1$ we have $y_{i'} = y_i = 1$ and thus $\overline{y}_i^{s_{w_J}} = 1 = y_i$. Similarly, for $i = 2p+1$ and $i' = 2p+2$ are similar only to each other and there we also obtain perfect prediction: $\overline{y}_i^{s_{w_J}} = 0 = y_i$. To conclude, $SSE(J) = MSE(J) = 0$. Thus, $AMSE(J, c) = MSE(J) + c|J| = ck = R$.

Conversely, assume that $J \subseteq M \equiv \{1, ..., m\}$ is such that $AMSE(J, c) \leq R$. We argue that we have to have $SSE(J) = MSE(J) = 0$. To see this, assume, to the contrary, that J does not provide a perfect fit. Thus, there exists i such that $\overline{y}_i^{s_{w_J}} \neq y_i$. As $y_i \in \{0, 1\}$ and $\overline{y}_i^{s_{w_J}}$ is a relative frequency in a bin of size no greater than n, the error $\left|\overline{y}_i^{s_{w_J}} - y_i\right|$ must be at least $\frac{1}{n}$. Therefore, $SSE(J) \geq \frac{1}{n^2}$ and $MSE(J) \geq \frac{1}{n^3}$. However, $R = ck \leq cm$ and as $c < \frac{1}{mn^3}$ as we have $cm < \frac{1}{n^3}$. Hence $MSE\left(J\right) \geq \frac{1}{n^3} > cm \geq R$, that is, $MSE\left(J\right) > R$ and $AMSE(J,c) > R$ follows, a contradiction.

It follows that, if J obtains a low enough AMSE ($AMSE(J,c) \leq R$), it obtains a perfect fit. This is possible only if within each J-bin the values of y_i's are constant. In particular, the observations $i = 2p+1$ and $i' = 2p+2$ (which, being identical are obviously in the same bin) are not similar to any other. That is, for every $i \leq 2p$ we must have $s_{w_J}(x_i, x_{2p+1}) = 0$. This, in turn, means that for every such i there is a $j \in J$ such that $x_i^j \neq x_{2p+1}^j$. But $x_{2p+1}^j = 0$ so this means that $x_i^j = 1$. Hence, for every $u \leq p$ there is a $j \in J$ such that $x_{2u}^j = z_{uj} = 1$, that is, $\left\{T_v\right\}_{v \in J}$ is a cover of P. It only remains to note that $AMSE(J,c) \leq R$ implies that $|J| \leq k$. □

Proof of Proposition 2:

The relevant SSE's are given by

$$SSE\left(J\right) = \frac{\left(N+k\right)\left(N+l\right)^2 + \left(N+l\right)\left(N+k\right)^2}{\left(2N+k+l-1\right)^2}$$

and, for $k+l > 1$,

$$SSE\left(J \cup \{j\}\right) = \frac{2N^3}{\left(2N-1\right)^2} + \frac{kl^2 + lk^2}{\left(k+l-1\right)^2}$$

where the first summand is the contribution of the $2N$ cases with $x_i^j = 0$ and the second – the $(k+l)$ cases with $x_i^j = 1$, and

$$SSE\left(J \cup \{j\}\right) = \frac{2N^3}{\left(2N-1\right)^2} + 0.25$$

if $k+l = 1$.

To see (i), fix N and l. As $k \to \infty$, we have

$$\begin{aligned} SSE\left(J\right) &= \frac{\left(N+k\right)\left(N+l\right)^2 + \left(N+l\right)\left(N+k\right)^2}{\left(2N+k+l-1\right)^2} \\ &= \frac{\left(N+l\right)k^2 + ...}{k^2 + ...} \to N+l \end{aligned}$$

while

$$\frac{2N^3}{(2N-1)^2} < N$$

for $N > 2$ and

$$\frac{kl^2 + lk^2}{(k+l-1)^2} \rightarrow_{k\rightarrow\infty} l$$

hence, $SSE(J) > SSE(J \cup \{j\})$ for all k large enough.

As for (ii), let $l = 0$. Consider first $k = 1$. We then have

$$SSE(J) < SSE(J \cup \{j\})$$

iff

$$\frac{(N+1)N^2 + N(N+1)^2}{4N^2} < \frac{2N^3}{(2N-1)^2} + \frac{1}{4}$$

which is equivalent to

$$2N^2 + 2N - 1 > 0$$

which for $N = 1$ is satisfied. Hence $k(N, 0) > 1$ is established.

Next consider $k > 1$. We have

$$SSE(J) - SSE(J \cup \{j\}) = \frac{(N+k)N^2 + N(N+k)^2}{(2N+k-1)^2} - \frac{2N^3}{(2N-1)^2}$$

We find that $SSE(J) > SSE(J \cup \{j\})$ would hold whenever

$$\frac{2N^3 + 3kN^2 + k^2N}{4N^2 - 4N(k-1) + 1} - \frac{2N^3}{4N^2 - 4N + 1} > 0$$

Calculations that are less than insightful show that the above holds whenever $(5k-1)N > 3$, which is the case under our assumptions. It follows that for every $k > 1$ inclusion of j in the set of predictors results in a lower SSE, and $k(N, 0) = 2$ is established.

For part (ii), for all $k \geq l = 1$ we have:

$$SSE(J) - SSE(J \cup \{j\}) = \frac{(N+k)(N+1)^2 + (N+1)(N+k)^2}{(2N+k)^2} - \left(\frac{2N^3}{(2N-1)^2} + \frac{1+k}{k}\right) \tag{9}$$

Evaluating the r.h.s. of expression 9 at $k = 1, 2, 3, 4$ and 5, one obtains, respectively:

$$-2N\frac{-3N+12N^3+1}{(2N-1)^2(2N+1)^2}$$
$$-\frac{1}{4}N\frac{2N+12N^2-1}{(N+1)(2N-1)^2}$$
$$-\frac{2}{3}N\frac{26N+28N^2+8N^3-9}{(2N+3)^2(2N-1)^2}$$
$$-\frac{1}{4}N\frac{42N+18N^2-13}{(N+2)^2(2N-1)^2}$$
$$\frac{2}{5}N\frac{-187N-32N^2+12N^3+55}{(2N+5)^2(2N-1)^2}$$

The first four expressions are negative for $N > 0$, while the last one is

positive, hence we can conclude that $k(N, 1) = 5$. □

Proof of Theorem 3:

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 $O(n^2 m)$ steps as in the proof of Theorem 2. 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 that it is NPC is basically the same as that of Theorem 2, and we use the same notation here. That is, we assume a given instance of SET-COVER: P, $r \geq 1$ subsets thereof, $T_1, ..., T_r \subseteq P$, and an integer k, with $P = \{1, ..., p\}$, $\cup_{i \leq r} T_i = P$, and the incidence matrix $z_{uv} \in \{0, 1\}$. We let $n = 2(p+1)$ and $m = r$, and, for $u \leq p$, $i = 2u - 1, 2u$ is given by $x_i^j = z'_{uj}, y_i = 1$ whereas for $i = 2p+1, 2p+2$, $x_i^j = 0$ and $y_i = 0$. We again set $c = (mn^3)^{-1}/2$ and $R = kc$. This construction can obviously be done in polynomial time. We claim that there is a cover of size k of P iff there is a vector of extended non-negative rationals w such that $AMSE(w, c) \leq R$.[18]

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. 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 $w = (w^j)_j$ ($w^j \in [0, \infty]$) 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, as in the proof of Theorem 2, 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 $\overline{y}_i^{s_w} < 1 = y_i$. In the proof of Theorem 2 this sufficed to bound the error $|\overline{y}_i^{s_w} - y_i|$ from below by $\frac{1}{n}$, as all observations with positive weights had the same weights. This is no longer the case here. However, the cases $2p+1, 2p+2$ obtain maximal similarity

[18] 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.

to i $(s(x_i, x_{2p+1}) = s(x_i, x_{2p+2}) = 1)$, because $x^j_{2p+1} = x^j_{2p+2} = x^j_i (= 0)$ for all j with $w^j > 0$. (It is possible that for other observations $l \leq 2p$ we have $s(x_i, x_{2p+1}) \in (0, 1)$, which was ruled out in the binary case. But the weights of these observations is evidently smaller than that of $2p + 1, 2p + 2$.) Thus we obtain (again) that the error $|\overline{y}^{s_w}_i - y_i|$ 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. $\square$

8 References

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

Akaike, H. (1974), “A New Look at the Statistical Model Identification”. *IEEE Transactions on Automatic Control* **19** (6), 716–723.

Anscombe, F. J. and R. J. Aumann (1963), “A Definition of Subjective Probability”, *The Annals of Mathematics and Statistics*, **34**: 199-205.

Aragones, E., I. Gilboa, A. Postlewaite, and D. Schmeidler (2005), “Fact-Free Learning”, *American Economic Review*, **95**: 1355-1368.

Argenziano, R. and I. Gilboa (2012), “History as a Coordination Device”, *Theory and Decision*, **73**: 501-512.

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler (2005), “Probabilities as Similarity-Weighted Frequencies”, *Econometrica*, **73**: 1125-1136.

Billot, A., I. Gilboa, and D. Schmeidler (2008), “Axiomatization of an Exponential Similarity Function”, Mathematical Social Sciences, **55**: 107-115.

Bray, M. (1982), “Learning, Estimation, and the Stability of Rational Expectations”, *Journal of Economic Theory*, **26**: 318-339.

de Finetti, B. (1931), Sul Significato Soggettivo della Probabilità, *Fundamenta Mathematicae*, **17**: 298-329.

——— (1937), “La Prevision: ses Lois Logiques, ses Sources Subjectives”, Annales de l’Institut Henri Poincare, **7**: 1-68.

Eilat, R. (2007), “Computational Tractability of Searching for Optimal Regularities”, working paper.

Fix, E. and J. Hodges (1951), “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

——— (1952), "Discriminatory Analysis: Small Sample Performance". Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Gilboa, I. and D. Schmeidler (1995), "Case-Based Decision Theory", *The Quarterly Journal of Economics*, **110**: 605-639.

——— (2001), *A Theory of Case-Based Decisions*, Cambridge: Cambridge University Press.

——— (2012), *Case-Based Predictions.* World Scientific Publishers, Economic Theory Series (Eric Maskin, Ed.), 2012.

Gilboa, I., O. Lieberman, and D. Schmeidler (2006), "Empirical Similarity", *Review of Economics and Statistics*, **88**: 433-444.

Hume, D. (1748), *An Enquiry Concerning Human Understanding.* Oxford: Clarendon Press.

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

Ramsey, F. P. (1926a), "Truth and Probability", in R. Braithwaite (ed.), (1931), *The Foundation of Mathematics and Other Logical Essays.* London: Routledge and Kegan.

Ramsey, F. P. (1926b), "Mathematical Logic", *Mathematical Gazette*, **13**: 185-194.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function", *Annals of Mathematical Statistics*, **27**: 832-837.

Savage, L. J. (1954), *The Foundations of Statistics.* New York: John Wiley and Sons. (Second addition in 1972, Dover)

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: John Wiley and Sons.

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

Steiner, J., and C. Stewart, C. (2008), "Contagion through Learning", *Theoretical Economics*, **3**: 431-458.