Learning What is Similar: Precedents and Equilibrium Selection*

Rossella Argenziano† and Itzhak Gilboa‡

March 2018

Abstract

We argue that a precedent is important not only because it changes the relative frequency of a certain event, making it positive rather than zero, but also because it changes the way that relative frequencies are weighed. Specifically, agents assess probabilities of future events based on past occurrences, where not all of these occurrences are deemed equally relevant. More similar cases are weighed more heavily than less similar ones. Importantly, the similarity function is also learnt from experience by "second-order induction". The model can explain why a single precedent affects beliefs above and beyond its effect on relative frequencies, as well as why it is easier to establish reputation at the outset than to re-establish it after having lost it. We also apply the model to equilibrium selection in a class of games dubbed "Statistical Games", suggesting the notion of Similarity-Nash equilibria, and illustrate the impact of precedents on the play of coordination games.

*Gilboa gratefully acknowledges ISF Grant 704/15 and the Investissements d'Avenir ANR -11- IDEX-0003 / Labex ECODEC No. ANR - 11-LABX-0047.

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

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

1 Introduction

1.1 Motivating Examples

1.1.1 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. This was clearly an important precedent: 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.

The importance of this single case does not seem to be fully captured by the change in the relative frequency of African-American presidents, 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 on the perceived likelihood of future presidents being African-American does not seem to be captured by the difference between 0:11 and 1:12. We suggest that the importance of the precedent set by Obama is partly explained by a process of "second-order induction". According to this view of learning, past data are used in two ways: through first-order induction, to estimate the probabilities of future events according to the relative frequency of similar events in the past, and through second-order induction, to learn what counts as "similar", hence relevant for prediction. Up to Obama's election, "race" was an important attribute in assessing the probability that a given candidate might be elected. But once the precedent of Obama was set, people who look at history may conclude that the race variable is not necessarily helpful in explaining past data and predicting future outcomes. By suggesting that the notion of similarity between cases is updated as new data are observed, second-order induction helps explain the dramatic importance of precedents.

1.1.2 The Fall of the Soviet Bloc

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.

It has been argued that similarity-weighted frequencies of past cases can be used to predict of the outcome of revolution attempts[1]. The case of Poland was an important precedent, which generated a "domino effect". We suggest that its importance didn't lie only in changing the relative frequency of successful revolutions, but also in changing the notion of which past revolution attempts were similar to current ones, hence relevant to predict their outcomes, via second-order induction. Specifically, the case of Poland was the first revolution attempt after the "Glasnost" policy was declared and implemented by the USSR. Pre-Glasnost attempts in Hungary in 1956 and in Czechoslovakia in 1968 had failed. In 1989, one might well wonder, has Glasnost made a difference? Is it a new era, where older cases of revolution attempts are no longer relevant to predict the outcome of a new one, or is it "Business as usual", and Glasnost doesn't change much more than, say, a leader's proper name, leaving pre-Glasnost cases relevant for prediction?

If the revolution attempt in Poland were to fail like the previous ones, it would seem that the variable "post-Glasnost" does not matter for prediction: with or without it, revolution attempts fail. As a result, second-order induction would suggest that the variable "post-Glasnost" be ignored, and the statistics would suggest zero successes out of 3 revolution attempts. By

[1]Revolution attempts can be modelled as coordination games, because the expected value from taking part in an uprising increases in the probability of it success, hence in the number of participants. (See, for example, Edmond, 2013). Steiner and Stewart, 2008, Argenziano and Gilboa, 2012, and Halaburda, Jullien, and Yehezkel, 2017 provide models in which similarity-weighted frequencies of past cases are used to form beliefs in coordination games.

contrast, because the revolution attempt in Poland succeeded, it had a double effect on the statistics. By first-order induction alone, it increased the frequency of successful revolutions from 0:2 to 1:3, which is still less than a half, and still leads to pessimistic predictions about future attempts. However, by second order induction, the post-Glasnost variable is learned to be important, because the frequency of success post-Glasnost, 1:1, differs dramatically from the pre-Glasnost frequency, 0:2 . Once this is taken into account, pre-Glasnost events are not as relevant for prediction as they used to be. If we consider the somewhat extreme view that post-Glasnost attempts are in a class apart, the relevant empirical frequency of success becomes 1:1 rather than 1:3. Correspondingly, other countries in the Soviet Bloc could be encouraged by this single precedent, and soon it wasn't single any more.

In our first motivating example (Obama's election), we find that a precedent makes a variable lose relevance: race used to be considered a variable with predictive power, restricting attention to sub databases defined by race. The single precedent was enough to suggest that race is unimportant, and a candidate's probability of success should be assessed based on other variables. In the second example (Poland's revolution) the opposite happened: a precedent introduced a new variable into similarity judgments. The single case of Poland convinced people that this is "a new ballgame", and that the relevant database to look at is the restricted one of post-Glasnost attempts. We seek to develop a theory that can capture both these examples, and examine its implications for some applications, most notably equilibrium selection in coordination games.

1.2 Belief Formation

How do agents form beliefs about the likelihood of future events? In many cases, the answer is within the realm of statistics. When evaluating the probability of a car theft, for example, one may rely on empirical frequencies, which provide natural estimators of probabilities when observations can

be viewed as realizations of i.i.d. random variables. In other problems, such as assessing the probability of developing a disease, more sophisticated techniques are used in statistics and machine learning, allowing for learning from cases that are not identical and for identifying patterns in the data. Thus, logistic regression, decision trees, non-parametric methods and many other techniques can be used to provide probabilistic assessments. However, there are many problems in which there are relatively few observations, and those that exist are rather different from each other. For example, in assessing the probability of success of a presidential candidate, past cases are clearly of relevance, but no two are similar enough to simply cite empirical frequencies. The focus of this paper is the belief generation process in these decision problems.

We consider a very simple model, according to which the probability of an event is taken to be its similarity-weighted relative frequency. Thus, the probability that a candidate will win the election is estimated by the proportion of cases in which similar candidates won elections, where more similar candidates are assigned higher weights than less similar ones. The determinant of similarity may include factors such as party affiliation, political platform, and experience, as well as gender, race, and age[2]. Our main point is that the *way* similarity of cases should be judged is itself learnt from the data. Whereas learning from past cases about the likelihood of future ones is referred to as *first-order induction*, learning the similarity function, namely, the way first-order induction should be conducted, is dubbed *second-order induction*.

Using similarity-weighted averages is an intuitive idea that appeared in statistics as "kernel methods" (Akaike, 1954, Rosenblatt, 1956, Parzen, 1962). Further, statistical methods also suggest finding the optimal bandwidth of the kernel function (Nadaraya,1964, Watson, 1964), which is concep-

[2]Clearly, this model is simplistic in many ways. For example, it does not allow for the identification of trends, as logistic regression would. Yet, it suffices for our purposes.

tually similar to the second-order induction studied here. Interestingly, very similar processes were also suggested in psychology. The notion of "exemplar learning" (see Shepard, 1957, Medin and Schaffer, 1978, and Nosofsky, 1984) suggests that, when people face a categorization problem, the probability they would choose a given category can be approximated by similarity-weighted frequencies. Further, it has also been shown that people learn the relative importance of different attributes in making their similarity judgments (Nosofsky, 1988, see Nosofsky, 2011 for a survey). Categorization in general, and optimal categorization in particular, has also been suggested by Fryer and Jackson (2008).

This paper is closer to Gilboa, Lieberman, and Schmeidler (GLS, 2006), who suggested the notion of learning the similarity function from the data, and referred to the optimal function as the "empirical similarity". While their paper can be viewed as suggesting a statistical technique, similar to the choice of an optimal bandwidth in kernel estimation, our focus in this paper is on the interpretation of the model as a description of the way people reason. Note that the psychological evidence cited above deals with learning a similarity function for the purpose of a categorization task, which is distinct from (and perhaps cognitively less demanding than) the estimation of probabilities. Yet, we find such learning to be rather intuitive. For example, a physician who has to estimate the probabilities of success of a medical procedure would rely on past data, and would use her experience to learn which medical variables are more important than others. Similarly, in our motivating example, a potential donor who tries to estimate a candidate's probability of winning would also look at past data, and use these data to learn how much weight each observation should be assigned.

Argenziano and Gilboa (2017) study a second-order induction model where the empirical similarity is computed by a leave-on-out cross-validation technique. The focus of that paper is on asymptotic results regarding the uniqueness of the empirical similarity function and the complexity of its com-

putation, in particular when the number of relevant variables can be rather large. By contrast, in this paper we consider the same model and study conditions under which a single variable – such as "race" or "post-Glasnost" in the examples above – will be included in the empirical similarity function. Abstracting away from the other variables, and focusing on binary variables throughout, we deal with a seemingly very simple problem, characterized by no more than four parameters. We provide some results about values of these parameters for which the similarity will, or will not, include a specific variable, and show that the model captures the intuitions explained in subsection 1.1.

1.3 Equilibrium Selection

A theory of belief formation might be a building block in a theory of equilibrium selection. Indeed, if we know how people form beliefs, we can predict that they best-respond to these beliefs, and if these best responses define an equilibrium, this equilibrium would be a more likely prediction than others. Indeed, the example of the collapse of the Soviet Bloc is naturally conceptualized as a sequence of coordination games, with one equilibrium describing a successful revolution and the other – no revolution attempt.

Embedding statistical learning in a theory of equilibrium selection raises two issues having to do with strategic considerations. First, if players in a game are aware of the fact that other players are also strategic, they will not predict their behavior as if it simply were a natural phenomenon; they would take into account other players' predictions, their predictions of others' predictions and so forth. Thus, there is a gap between beliefs formed using statistical and strategic reasoning. The former ignores the fact that other players also learn from data, while the latter allows data to be completely ignored. Second, there is also inter-period strategic reasoning. If players use past data to make predictions and decisions, a current choice might have to take into account its possible effects on the statistical learning of others in

the future.

To deal with the first problem, we suggest to merge statistical and strategic reasoning: statistics is used to generate initial beliefs p about the play of the game (shared by all players), and these beliefs are used to compute best responses. If these result in an equilibrium, we suggest it as a natural candidate for the prediction of the way the game will be played. That is, purely statistical, non-strategic reasoning is used to suggest naive beliefs, and these are fed into strategic reasoning. We only use these initial beliefs if the best responses to them are also best responses to themselves, that is, as an equilibrium selection device. In other words, the naive, non-strategic statistics are used as focal points for the game.

Inter-period strategic considerations may be very important in some setups, but not in all. Polish citizens who had to decide whether to join the revolution attempt in 1989 are unlikely to have put much weight on the impact of their decision on a future revolution in Czechoslovakia. We suggest to model these consecutive revolution attempts as a "statistical game". A statistical game is defined as a sequence of games played by disjoint sets of players, with no direct strategic considerations across games played in different periods. However, each game starts with a draw of a set of variables $x = (x^1, ..., x^m)$, and, after the players make their moves, a realization of a variable y. Further, the payoff of each player depends on the realization of x, on the player's own move, and on y (but not on others' moves given y). Thus, it makes sense for each player to try to predict y based on its past realizations in similar games. We assume that this prediction is done according to similarity-weighted frequencies employing an empirical similarity function. This process offers initial beliefs that can be fed into the strategic reasoning process. Equilibria that can be justified by this process are dubbed *Similarity-Nash* equilibria. We analyze a simple coordination game (modeled as a statistical game) and show that a single precedent, such as a successful revolution in Poland, defines a unique Similarity-Nash equilibrium

corresponding to the intuition described above. In the coordination game we consider the variables $x = (x^1, ..., x^m)$ are payoff-neutral and can be viewed as "sunspots" (Cass and Shell, 1983) that are commonly observed and used for coordination in the game. As such, our theory of finding the optimal similarity function can be viewed as a theory of sunspot selection.

The rest of the paper is organized as follows. Section 2 presents the statistical model and the notion of empirical similarity. Section 2.2 focuses on a single variable and analyzes the importance of precedences from the perspective of the empirical similarity model. The analysis is extended beyond precedents to general problems, asking under which condition the variable in question will be included in an empirical similarity function. In particular, the results show why, in this model, it is easier to establish reputation than to re-establish it. Section 3 defines Statistical Games and Similarity-Nash equilibria formally and applies the analysis to an example of equilibrium selection in coordination games. Finally, Section 4 concludes with a general discussion.

2 Second-Order Induction in Prediction Problems

2.1 Case-Based Beliefs and Second-Order Induction

A binary variable $y \in \{0, 1\}$ is to be predicted based on other binary variables, $x^1, ..., x^m \in \{0, 1\}$. We assume that there are n observations of the values of $x = (x^1, ..., x^m) \in X \equiv \{0, 1\}^m$ and of the corresponding y values. Given a new value for the x's, an agent attempts to predict the value of y. Observations will be denoted by subscripts, so that observation i is (x_i, y_i) where $x_i = (x_i^1, ..., x_i^m) \in X$ and $y_i \in \{0, 1\}$. A new data point x_p is given, and the agent attempts to predict y_p.

We assume that prediction is made by a similarity function $s : X \times X \to \mathbb{R}_+$, such that the probability that $y_p = 1$ is estimated by the similarity-

weighted empirical frequency

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

if $\sum_{i \leq n} s(x_i, x_p) > 0$ and $\overline{y}_p^s = 0.5$ otherwise.[3]

In this paper we focus on a simple model, according to which the similarity function takes values in $\{0, 1\}$. Further, we assume that for any given similarity function, each variable either counts as relevant for prediction, or as irrelevant.[4] Thus, for a subset of predictors, $J \subset M \equiv \{1, ..., m\}$, let the associated similarity function be:

$$s_J(x_i, x_p) = \prod_{j \in J} \mathbf{1}_{\left\{x_i^j = x_p^j\right\}} \tag{2}$$

In other words, the similarity of two vectors is 1 iff they are identical on the set of relevant variables, J. Clearly, the relation "having similarity 1" is an equivalence relation.

We introduce the notion of *second-order induction* to capture the idea that in order to obtain more accurate predictions, agents choose the similarity function that best fits the data. We define the "empirical similarity" as a similarity function that, had it been used to predict the existing data points, where each is estimated based on the others, would have performed best. In particular, we consider a leave-one-out cross-validation technique as a model of the process people implicitly undergo in learning similarity from data. Formally, for each subset of predictors, $J \subset M$, let

$$\overline{y}_i^J = \frac{\sum_{r \neq i} s_J(x_r, x_i) y_i}{\sum_{r \neq i} s_J(x_r, x_i)}$$

[3]This formula can be extended to the case of more than two possible values for the predictors x^j and for y in a straightforward manner.

[4]Argenziano and Gilboa (2017) deal with this binary model as well as with a model in which both the variables and the similarity function are continuous. They focus on asymptotic analysis, and find similar results for the two models.

and consider the sum of squared errors,

$$SSE\left(J\right)=\sum_{i=1}^{n}\left(\overline{y}_{i}^{J}-y_{i}\right)^{2}$$

A function s_J such that $J \in \arg\min SSE(J)$ is *an empirical similarity function.*

Observe that the empirical similarity need not be unique. To consider the most trivial case, suppose that a variable x^j is constant in the database. In this case, $SSE\left(J\right)=SSE\left(J\cup\left\{j\right\}\right)$ for any $J\subset M$. By convention, we may decide to drop such a variable (j), implicitly assuming that handling a variable incurs some memory and computation costs that are assumed away in this paper. However, there could be more interesting examples of non-uniqueness. See Argenziano and Gilboa (2017) for details.

2.2 When is a Variable Relevant?

The focus of our analysis is the question of a variable's relevance for prediction. Formally, given a set of predictors, $J\subset M$ and $j\notin J$, we are interested in the comparison of $SSE\left(J\right)$ and $SSE\left(J\cup\left\{j\right\}\right)$. If $SSE\left(J\right)>SSE\left(J\cup\left\{j\right\}\right)$, then the inclusion of the variable j provides a better fit to the data. We then assume that people would take this variable into account when assessing the probability that $y=1$ in the next observation. If, by contrast, $SSE\left(J\right)<SSE\left(J\cup\left\{j\right\}\right)$, the addition of the variable j to the similarity function results in higher errors, and we assume that the variable will be ignored by people who assess this probability. The reason that more variables can result in worse predictions is related to "the curse of dimensionality": a set of predictors J splits the database into sub-databases with identical $\left(x^{l}\right)_{l\in J}$ values. A new variable splits each of these sub-databases into smaller ones, so that their number grows exponentially in $|J|$. When there are too few observations in a sub-database, the prediction error can

grow[5].

Whether a set of predictors J will perform better by the addition of a variable $j \notin J$ depends mostly on how much information the latter carries about y, *given the variables* J. In general, this information need not be summarized by simple correlations or regularities. It is possible that for some $\left(x^l\right)_{l\in J}$ values of the variables in J, $x^j = 1$ makes $y = 1$ more likely, and vice versa for other $\left(x^l\right)_{l\in J}$ values. While such cases are theoretically interesting and important, they seem to be more involved than our motivating examples.[6] We wish to focus attention on simple cases, in which, should a variable be included, it is relatively clear what predictions it induces. We therefore assume $J = \varnothing$ and address the question of whether a variable x^j should be included in the similarity function.

Intuitively, the question is about the difference in the proportion of cases with $y = 1$ (vs. $y = 0$) in the two sub-databases, one with $x^j = 1$, and its complement, with $x^j = 0$. If the proportion is the same, there is no predictive power to be gained from splitting the database according to x^j. If, by contrast, the proportion is different, then x^j provides statistical information about y. Whether the additional statistical information is worth splitting the database into two smaller sub-databases would depend on the sizes of the sub-databases obtained, due to the curse of dimensionality discussed above.

The n points in the database are divided into four types, according to the values of x^j and of y. Let the number of cases of each type be given by the following case-frequency matrix:

# of cases	$x^j = 0$	$x^j = 1$
$y = 0$	L	l
$y = 1$	W	w

[5] Note that this reason is distinct from overfitting, which may be yet another reason to prefer small sets of predictors.

[6] Again, see Argenziano and Gilboa (2017) for discussion of the problem in the general case, including problems having to do with computational complexity.

We are interested in the sign of

$$\Delta(L, W, l, w) \equiv SSE(\{j\}) - SSE(\varnothing)$$

where $\Delta(L, W, l, w) > 0$ implies that the variable j is not included in the empirical similarity function, whereas $\Delta(L, W, l, w) < 0$ implies that it is. Clearly, $\Delta(L, W, l, w) = \Delta(W, L, w, l)$ and $\Delta(L, W, l, w) = \Delta(l, w, L, W)$, as the SSE calculations do not change if we switch between 0 and 1 either for a predictor x^j or for the predicted variable y.

We assume that there is a non-trivial history in which $x^j = 0$. Specifically, we assume throughout that $L, W > 2$. This assumption means that (i) the database contains a non-trivial number of cases overall, and that (ii) the prediction of the variable in question, y, is a non-trivial task: there are a few (at least three) cases with $y = 0$ as well as with $y = 1$.

Our focus in this paper is on databases for which the number of cases with $x^j = 1$ is small. We wish to study the change of beliefs when a new event occurs – such as the election of an atypical candidate for the presidency, or the behavior of a new agent who has no history, and so forth. For these cases we will think of w and l as small (and sometimes zero). Databases with $w = l = 0$ will be of special interest. They can be interpreted in two ways, between which our model does not attempt to distinguish: first, it is possible that all relevant agents are aware of the variable x^j, and they notice that $x^j = 1$ has never been observed. Second, they might be situations in which the variable x^j hasn't really occurred to anyone because it has never been observed. For example, in the application of the model to the study of reputation, the variable in question will be an agent's proper name, and agents were probably not aware of the variable before a person with that proper name appears on stage. We do not attempt to distinguish between the two interpretations, and do not need to for the sake of the model.

2.2.1 Simple Regularities

The first result we establish is that, if there are sufficiently robust regularities in the database, the empirical similarity will spot them. In particular, suppose that the database contains at least two cases with $x^j = 1$, and *all* such cases have the same y value. Then, we prove that the variable j will be included in the empirical similarity function, as it will be perceived to be of predictive power. Formally,

Proposition 1 For any (L, W), and any $l, w > 1$, we have

$$\Delta(L, W, 0, w) \ , \ \Delta(L, W, l, 0) < 0.$$

Recall that we assume that $L, W > 2$, so that the sub-database for which $x^j = 0$ does not suggest a clear regularity about y. By contrast, in the sub-database for which $x^j = 1$, y is constant. If there are at least two cases in this sub-database, second-order induction will "identify" the regularity and include the variable j in the empirical similarity function. Proposition 1 is rather intuitive and turns out to be very simple to prove. Yet, it is important because it shows that, if case-based predictions are allowed to use second-order induction, they will not miss simple regularities in the data.

The parameter values $w = 1, l = 0$ (or vice versa, $w = 0, l = 1$) are not covered by Proposition 1 but they are particularly interesting. They correspond to new realities, where $x^j = 1$ has never been observed before. Our next result shows that, when a case with $x^j = 1$ is observed for the first time, the variable j will be included in the empirical similarity if and only if the corresponding y value was *the less frequent value* in the rest of the database. Formally,

Proposition 2 If $W < L$, $\Delta(L, W, 0, 1) < 0$ and $\Delta(L, W, 1, 0) > 0$. Symmetrically, if $W > L$, $\Delta(L, W, 0, 1) > 0$ and $\Delta(L, W, 1, 0) < 0$. Finally, $\Delta(W, W, 1, 0), \Delta(W, W, 0, 1) > 0$.

We find this result rather intuitive: when no cases with $x^j = 1$ were ever observed ($w = l = 0$), there is no real meaning to the variable x^j: it is always 0 and can be ignored.[7] When the first case with $x^j = 1$ pops up, one is led to ask, is this new feature useful? Should I make a note of the fact that the new case had this new feature, or should I better dismiss it as noise? For example, suppose that one is watching horse races, and classifies horses into "very fast" ($y = 1$) or "regular" ($y = 0$), where the majority of the horses are "regular". At some point one observes, for the very first time ever, a green horse. Stunning as this phenomenon is, the unusual color might not be informative. Proposition 2 says that, *if* the green horse turns out to be very fast, the next time a green horse will show up its color would be noticed. By contrast, if the conspicuously colored horse turns out to be regular, the special feature will be dismissed.

2.2.2 Losing Relevance through a Precedent

Our first two propositions investigated databases in which a simple regularity holds: $x^j = 1$ implies a specific value for y in every single observation in the database. We now turn to the case in which no such rule holds, and for $x^j = 1$ both cases with $y = 0$ and with $y = 1$ have been observed. When should the variable be included in the empirical similarity?

We first study the impact of a single precedent. Proposition 1 established that if there are at least two cases in which $x^j = 1$, and they all have the same outcome, then the variable has enough predictive power to be included in the empirical similarity. Proposition 3 shows that a single case will reverse this result, unless the number of cases that established the regularity is sufficiently large:

Proposition 3 For every $0 < l \leq \left\lfloor \frac{L}{W} \right\rfloor + 1$, we have $\Delta(L, W, l, 1) > 0$.

[7] As mentioned above, in this case (where we have, in particular, $\Delta(K, L, 0, 0) = 0$), we assume that j is not included in the optimal set of predictors.

We interpret Proposition 3 as capturing the way that a single precedent makes a variable lose importance. Consider our motivating example, namely, the election of President Obama. We focus on the variable x^j denoting race, where $x^j = 1$ means that the candidate is African-American. Assume that the database in a typical voter's mind includes all cases of people who ran for the Democratic or Republican parties' nomination since WWII, where $y = 1$ stands for "was elected President". The vast majority of them were white, namely, had $x^j = 0$. Of these, W won and L lost, where L is significantly larger than W. ($L/W \approx 10$ would seem like a reasonable assumption about voter's perception of a typical campaign.) On top of these white candidates, there are also some attempts made by African-American candidates, but all of those failed. Assume that the number of these attempts is $l < 10$.[8]

By Proposition 1, given zero successes by African-American candidates, $w = 0$, and at least two failures, $l > 2$, the variable "race" has predictive power and the empirical similarity function takes it into account. That is, the probability of a successful campaign by an African-American candidate would be estimated to be the relative frequency of successes in the sub-database of African-Americans, namely, 0. This is the sense in which our model captures the regularity "No African-American was ever elected president". However, after Obama's election, the number of successes changed to $w = 1$. With $l < 10 \leq \left\lfloor \frac{L}{W} \right\rfloor + 1$, Proposition 3 implies that $\Delta(L, W, l, 1) > 0$. That is, the single case of Obama suffices to change the similarity function and drop "race" from the list of important variables. The probability of success of a future campaign of an African-American candidate would be judged according to other variables alone.[9]

[8] Relatively well-known campaigns are those of Shirley Chisholm (in 1972) and Jesse Jackson (in 1984, 1988). There were several more, and our search came up with 6 such campaigns. As far as a typical voter's memory is concerned, $l < 10$ seems to be a reasonable assumption.

[9] This probability is still relatively low, given that there are many more candidates who lost than candidates who won. The point, however, is that an African-American candidate would have the same perceived probability of success as a white candidate, while this was

Notice that the result need not apply if l is larger than $\left\lfloor \frac{L}{W} \right\rfloor + 1$. That is, if the proportion of successes among African-Americans given the Obama precedent, $1/(l+1)$, were still smaller than the corresponding proportion among the rest, $W/(L+W)$, the variable "race" could still be part of the similarity function. Indeed, if the case of Obama were to change the proportion of African-American successes from 0 to, say, 0.001, it would still be much lower than the proportion of successes among the other (white candidate) campaigns, and one would expect that the two populations would still be perceived as different for prediction purposes. (Proposition 4 below implies that this is indeed the case if l is sufficiently large.) The reason that a single precedent could change the similarity function so dramatically is that the number of failed attempts among African Americans was not that large.

2.2.3 Gaining Relevance

Consider an agent who's new to an economic or political scene, and who wishes to bring about a belief in a certain "success" outcome, $y = 1$, where this outcome used to be the exception rather than the rule. Thus, statistical analysis that takes into account all of history would suggest that $y = 0$ is more likely than $y = 1$, and our agent tries to establish a reputation that would dissociate her from past experiences. For example, the agent may be a new dean who aims to enforce regulations more strictly than her predecessors, or a central banker who intends to curb inflation. Let x^j be the indicator variable of the agent's proper name, so that, starting with a clean slate, there are no cases with $x^j = 1$, and $w = l = 0$. However, past cases (for which $x^j = 0$) have $W < L$, and it is this tradition of failures that the agent wishes to break away from.

Proposition 2 suggests that the new agent would have to invest an effort in establishing $y = 1$ once in order to establish her reputation: with $W < L$, $\Delta(L, W, 0, 1) < 0$, and x^j would already enter the empirical similarity

not the case before the precedent.

function. In our example, a single "success" can suffice for the dean to convey the message that "the rules have changed".

However, what will happen if the dean fails to enforce the rules at the beginning of her tenure? Intuition suggests that in that case she could still establish her reputation later on, but that this would become more costly. To analyze this scenario, we wish to study the function $\Delta(L, W, l, w)$ where both l and w are allowed to be beyond 1.

Proposition 4 studies the behavior of $\Delta(L, W, l, w)$ as a function of each of its two last arguments. In light of the symmetry $\Delta(L, W, l, w) = \Delta(W, L, w, l)$, it suffices to study one of them, which we take to be w for simplicity of notation. We start from a scenario in which the sub-database with $x^j = 1$ has, up to integrality constraint, the same ratio of cases with $y = 0$ and $y = 1$ as the sub-database with $x^j = 0$. If this ratio is precisely the same, that is, $\frac{w}{l} = \frac{W}{L}$, then x^j is irrelevant for predicting y in future cases, and we would expect j to be excluded from the optimal similarity function, that is, $\Delta(L, W, l, w) > 0$. It turns out that this is the case also if w is only known to be the closest integer to $\frac{lW}{L}$, or one above it. (Part (i) of the Proposition 4.) Suppose that we now increase w. We find that this improves the performance of the similarity function that includes the variable, up to a point where it outperforms the similarity function that does not include it. That is, if w/l is sufficiently larger than W/L (where the exact values of the parameters matter, and not only their ratios), then $\Delta(L, W, l, w) < 0$ (Part(ii)). As could be expected, the minimum $w^* > \frac{lW}{L}$ for which this inequality holds increases in the number of cases with the opposite outcome, l (Part (iii)). Moreover, up to details of integrality constraints, the number of *additional* cases needed to get to this minimum $(w^* - w)$ is also non-decreasing in l (Part (iv)).

Formally, let $[\ ] : \mathbb{R} \to \mathbb{Z}$ be the nearest integer function, selecting the truncation in case of a tie. (That is, for all $x \in R$ and $z \in Z$, we have $[x] = z$ if $x = z + \varepsilon$ and $\varepsilon \in [-0.5, 0.5)$.) We prove the following:

Proposition 4 Let $L, W, l, w > 0$ be any four integers such that $L, W > 2$, $l, w > 0$, and $w = \left[\frac{lW}{L}\right]$. The following hold:

(i) $\Delta(L, W, l, w)$, $\Delta(L, W, l, w+1) > 0$.

(ii) There exists an integer $w^*(L, W, l) \geq w+2$ such that, for every $q \geq w$,

$$\begin{aligned} q &< w^*(L, W, l) \quad \Rightarrow \quad \Delta(L, W, l, q) \geq 0 \\ q &\geq w^*(L, W, l) \quad \Rightarrow \quad \Delta(L, W, l, q) < 0 \end{aligned}$$

(Clearly, if such an integer exists it is unique.)

(iii) $w^*(L, W, l)$ is non-decreasing in l.

(iv) If W/L is an integer, $(w^*(L, W, l) - w)$ is non-decreasing in l.

Thus, our model captures the fact that it is harder to re-establish reputation than to establish it at the outset. By Proposition 2, if $W < L$ and $l = 0$ a single success ($w = 1$) suffices to establish reputation. By Proposition 4, with $l = 1$ at least three such cases would be needed (parts (i)-(ii)). More generally, for any number of adverse outcomes $l > 0$ there is a sufficiently large number of successes w that would eventually make one's proper name an important variable (part (ii)), but the additional number of successes required increases (part (iii)), and it does so more than proportionally, up to integrality constraints (part (iv)). One does get a second chance to make a first impression, but it becomes costlier.

3 Statistical Games

We now generalize the prediction problems discussed in section 2 to allow for strategic interactions. A *statistical game* G^* is a (finite or infinite) sequence of period games $(G_i)_{i \geq 1}$. The game G_i has a finite and non-empty set of players H_i, where the H_i's are pairwise disjoint. Game G_i is played in three stages, as follows. First, (Stage 1) Nature moves and determines the values of m binary variables, $x_i = (x_i^1, ..., x_i^m) \in \{0, 1\}^m$. Then, (Stage 2) all the players observe x_i and make simultaneous moves: player $h \in H_i$ selects an

action $a^h \in A_i^h$ (where A_i^h is non-empty and finite). Finally, (Stage 3) Nature selects a value for a variable $y_i \in \{0,1\}$ and the game ends. The payoff for player $h \in H_i$ is a function of $\left(x_i, a^h, y_i\right)$. That is, a player's payoff depends on the others' moves only to the extent that these affect the outcome y_i. (For example, in the revolution game we present below, a player's payoff depends on whether a revolution attempt succeeds or not, as well as on her own choice of supporting it, but, given y_i, it is independent of the choices of the other players.) In other words, having observed x_i, y_i is a sufficient statistic for the strategic aspect of the game. We also assume that, at the beginning of period i, all the players in H_i observe the entire history of characteristics and outcomes of past games, $\left((x_r, y_r)\right)_{r<i}$ but not the actions that were taken in them.[10]

Statistical games span a gamut of social interactions that involve learning. On the one extreme, one may consider pure prediction problems like those in section 2, where, at period i, a predictor is asked to guess the value of $y_i \in \{0,1\}$ given the value of $x_i \in \{0,1\}^m$ and the history $\left((x_r, y_r)\right)_{r<i}$. This is a special case of a statistical game in which there is no strategic interaction whatsoever. We may think of the predictor at time i as the single player h in H_i, with a set of actions $A_i^h = \{0,1\}$, whose payoff function is the indicator of a correct guess. On the other extreme, statistical games may suppress the learning aspect and focus on the strategic one. For example, if there are no predictors ($m = 0$), the only thing that a player needs to consider is the distribution of y_i. This may capture coordination games in which the only role of history is to serve as a coordination device.

The notion of a statistical game, as well as the solution concept we suggest for such games below, are compatible with several sets of implicit assumptions about the players' information. At a minimal level, the players may not know the distribution of y_i given x_i , and they use the empirical similarity in

[10] Clearly, this assumption is important even though we already assumed that y_r is a sufficient statistic for payoffs at stage $r < i$. For example, it does not allow a player in stage i to follow a strategy that is a function of the move of another player in stage $r < i$.

order to estimate it. This is the most natural interpretation if we consider a non-strategic environment, such as a one-person prediction problem. Alternatively, one may adopt the standard (implicit) assumption in game theory, namely, that the game G^* is commonly known among its players. This would imply that players know the distribution according to which Nature chooses x_i, given past history, $((x_r, y_r))_{r<i}$, as well as the conditional distribution of y_i (given (i) the history $((x_r, y_r))_{r<i}$; (ii) the current x_i; and (iii) all players' moves). Importantly, the prediction of y_i based on the realization of x_i then becomes a prediction about the players' moves. For example, in a revolution game all players might know what it would take for a revolution to succeed $(y_i = 1)$, in terms of the players' choices. The belief about a revolution succeeding induces a belief about what the other players are about to do.

The assumption that the sets of players H_i are pairwise disjoint implies that equilibria of G^* are basically selections of equilibria in period games G_i, each defined by a realization of x_i, for each i and each $x_i \in \{0,1\}^m$ (that occurs with positive probability). That is, at an equilibrium of G^*, players in G_i have to choose best responses to the others' moves in that game, as they have no future to worry about. Conversely, a selection of an equilibrium in each period game G_i (for each possible realization of x_i) yields an equilibrium of G^*, because players of different periods' games cannot coordinate deviations from the equilibrium path. Thus, the structure of G^* guarantees that all G_i's are strategically independent games.

Note that the games G_i are unrelated to each other apart from the information about the variables $((x_i, y_i))$. They have disjoint sets of players, and may have completely unrelated sets of acts and payoff functions. The only feature that relates them is the fact that in each game there is a realization of x_i (before players choose their moves), and a realization of y_i (after they do).[11]

[11] We implicitly assume that all the players encode information in the same way and that they agree on the meaning of statements such as "$x_i^j = 0$" or "$y_i = 1$". If, for instance, different players think of a given case as a "success" $(y_i = 1)$ and others – as

3.1 Similarity-Nash Equilibria

In this sub-section, we introduce an equilibrium notion for statistical games that incorporates the notion of second-order induction. We assume that players use the information available about past games to form initial beliefs about the outcome of the current one, and consider equilibria in which players best-respond to these initial beliefs.

Let there be a given a statistical game $G^* = (G_i)_{i\geq 1}$ with variables $x = (x^1, ..., x^m) \in \{0,1\}^m$ and $y \in \{0,1\}$. Consider game G_i. Given Nature's move in Stage 1, x_i is observed. Using the database $((x_r, y_r))_{r<i}$ one obtains an empirical similarity function $s_i = s_J$ such $J \in \arg\min SSE(J)$ with

$$SSE(J) = \sum_{r<i} \left(\overline{y}_r^{s_J} - y_r\right)^2$$

This function defines a probability distribution for y_i, denoted p_{s_i}. Specifically,

$$\begin{aligned} p_{s_i}(y_i = 1) &= \overline{y}_i^{s_i} \\ p_{s_i}(y_i = 0) &= 1 - \overline{y}_i^{s_i} \end{aligned} \tag{3}$$

which we take to represent the beliefs of each player h in G_i about y_i, if she were to ignore strategic considerations completely.

Note that a strategy for player h in G_i, $\mathbf{a}^h$, maps all histories of the form $\left(((x_r, y_r))_{r<i}, x_i\right)$ into A_i^h. Such a strategy is a *best response to* p_{s_i} if $\mathbf{a}^h\left(((x_r, y_r))_{r<i}, x_i\right) \in A_i^h$ maximizes player h's payoff in G_i given x_i and the belief p_{s_i} about y_i. (Recall that h's payoff only depends on $\left(x_i, a^h, y_i\right)$, so that, given knowledge of x_i and beliefs over y_i, the argmax of h's expected payoff is well-defined.) Strategies $\left(\mathbf{a}^h\right)_h$ that are best responses to p_{s_i} and *also happen to be* best responses to themselves (that is, $\mathbf{a}^h$ is a best response to $\left(\mathbf{a}^{h'}\right)_{h'\neq h}$ for all h) are called *Similarity-Nash equilibrium.*

a "failure" ($y_i = 0$), without a 1-1 mapping between the diffenet languages they use, we cannot assume a common process of statistical learning.

Similarity-Nash equilibria seem natural under a variety of assumptions about the players' information and strategic sophistication. For example, if players do not engage in too involved strategic reasoning, they may be interested only in the bottom line captured by y_i, best-respond to its distribution and play the equilibrium strategies. Alternatively, they may use the estimate of y_i as an initial conjecture and then apply strategic reasoning along the lines of Level-K reasoning, where Similarity-Nash equilibria result from Level-1 reasoning that already results in an equilibrium. Further, one may implicitly assume that the players are sophisticated enough to understand the entire model, and they realize that choosing a way to reason about the game can be viewed as a strategic choice in a "reasoning game". Such a game may be a coordination game, and if all players reason in a given way, it is a best response for each to follow the same mode of reasoning. If Similarity-Nash equilibria exist in the actual game, the way of reasoning we offer is an equilibrium in the implicit reasoning game.

As our focus in this paper is second-order induction, we define Similarity-Nash equilibria relative to the initial beliefs p_{s_i}, namely, the empirical similarity relative frequencies. However, one could use other initial beliefs as the statistical starting point used for strategic reasoning. Specifically, one can define the initial beliefs by first-order induction, that is, using an exogenously given similarity function to provide the statistical reasoning. This is basically the equilibrium selection process assumed in Steiner and Stewart (2008) and in Argenziano and Gilboa (2012).[12]

Since we consider only binary variables y_i, it is convenient to consider games G_i with two strict (and thus pure) Nash equilibria for any possible

[12]Both papers study cased-based reasoning in a class of complete information normal form coordination games. Games differ by one payoff-relevant parameter, and the similarity between two games is a function of the difference between the values of this parameter in the two games. Myopic players play a new game in each period and assess the expected payoff of each action by its expected payoff, where the beliefs over the other players' choices are given by similarity-weighted frequencies.

x_i.[13] Similarity-Nash equilibria are suggested as a criterion for equilibrium selection between these equilibria. Specifically, each equilibrium has a set of beliefs such that the equilibrium strategies are the unique best responses to any beliefs in this set. Harsanyi and Selten's (1988) notion of risk dominant equilibria is based on the size of maximal such sets. Similarity-Nash equilibria ignore the size of these sets and focus on the value of the statistical estimate p_{s_i}. In a sense, Similarity-Nash equilibria can be viewed as replacing a uniform distribution over players' moves by statistical learning, which is possible when the game is embedded in a history of other games. The analogy to risk dominant equilibria is stronger when all stage games have the same number of players, the same set of moves and the same payoff function for each player. Statistical games allow more freedom in the statistical learning procedure, where only the (x_i, y_i) relate the games played in different stages.

One can also view Similarity-Nash equilibria as a possible formalization of Schelling's (1960) focal points: one way in which an equilibrium can be focal is that it has been played in the past. Thus, relative frequency offers a natural criterion for selection of an equilibrium in a game that is being played repeatedly, and Similarity-Nash equilibria focus on the relative frequency according to the empirical similarity. Similarity-Nash equilibria are also defined when the games G_i differ from each other, as long as the variables (x_i, y_i) relate them in a meaningful way. For example, y_i might indicate whether a Pareto-dominating equilibrium has been played in the past, and thus the model can capture Pareto-domination as a focal point, allowing statistical learning across very different games.

In Section 4 we discuss a generalization of Similarity-Nash equilibria that allows an iterative process of best-response reasoning, starting with the statistical estimate p_{s_i} and leading to an equilibrium of the stage game.

[13] When more strict equilibria are considered, it is natural to extend the analysis to y_i that can assume at least as many values as there are equilibria.

3.2 Example of Equilibrium Selection in a Coordination Game

We consider here a simple example of a sequence of revolution games played by disjoint populations. At period i game G_i is played, describing a potential revolution attempt in a new country i. The players H_i are citizens of country i. Each citizen h observes the realization of x, x_i, and has to decide whether to join the revolution attempt, $a^h = 1$, or not, $a^h = 0$. As a result of these choices, the revolution succeeds, $y_i = 1$, or fails, $y_i = 0$. Assume that, irrespective of the values of x_i, the revolution succeeds (Nature chooses $y_i = 1$) with probability $f(\alpha) \in [0,1]$ where α is the proportion of players (in H_i) that chose $a^h = 1$. Further, we assume that $f(\alpha)$ is increasing, with

$$\begin{aligned} f(0) &= \varepsilon \\ f(1) &= 1-\varepsilon \end{aligned}$$

for $\varepsilon \in (0, 0.5)$. The assumptions that $f(0) > 0$ and $f(1) < 1$ reflect the fact that the model is not expected to capture all the relevant factors, and allow us to assume various histories in a way that is compatible with the model.

Let the payoff of Player h be determined only by her choice and the success of the revolution:

Payoff to h	$y_i = 1$	$y_i = 0$
$a^h = 1$	1	0
$a^h = 0$	0	1

Thus, a player's best response is to join the revolution attempt if and only if she thinks it is more likely to succeed than to fail.

We wish to study a single variable x^j, such as "post-Glasnost" in Example 1.1.2 and ask when it will be used in the empirical similarity function, that is, when will it be a sunspot, given a fixed set J of other variables. To simply matters, assume that $m = 1$ and the question is whether x^1 is used for prediction or not. Thus, the history $\{(x_r, y_r) \mid r < i\}$ is summarized in four non-negative integers (L, W, l, w) as in Section 2.2. We allow history

to contain revolution attempts against other regimes as well, some of which have been successful. However, we assume that there are more unsuccessful than successful attempts in the database.

We can now state

Corollary 1 *Assume that $L > W > 2$. Then, at any Similarity-Nash equilibrium:*

(i) If $w = 0$ and $l = 1$, $a^h = 0$ for all $h \in H_i$;

(ii) If $w = 1$ and $l = 0$, $a^h = 1$ for all $h \in H_i$.

Recall that, before Glasnost (for $x^1 = 0$), most revolution attempts failed ($L > W$). We consider the first post-Glasnost attempt and apply Proposition 1. Should the revolution attempt fail ($w = 0$ and $l = 1$), the variable x^1 would be deemed irrelevant (by $\Delta(L, W, 1, 0) > 0$), and the probability of a revolution succeeding would be estimated by $W/(L+1) < 0.5$. Thus the best response of each player would be $a^h = 0$ and this is an equilibrium.

By contrast, if the revolution attempt succeeds (which has a positive probability even if no player chooses $a^h = 1$), then we're in case (ii), $w = 1$ and $l = 0$. Then the proposition states that $\Delta(L, W, 0, 1) < 0$ and thus x^1 will be part of the empirical similarity function. That is, the post-Glasnost period would be considered a new era, and older cases would not factor into the statistics. In the post-Glasnost sub-database the proportion of successes is $w/(w+l)$, that is 100%. The probability of a revolution success would then be estimated by $w/(w+l) = 1 > 0.5$. Thus the best response of each player would be $a^h = 1$ and this, again, is an equilibrium.

4 Discussion

4.1 Additional Examples

4.1.1 Example: The Collapse of the Soviet Bloc Revisited

The collapse of the Soviet Bloc involved more than one variable. While the Soviet Bloc fell apart, the USSR remained a unified state. Despite the fact that the USSR consisted of fifteen republics, some of which contained ethnic majorities that seemed unhappy with Russian domination, for two more years there were no revolution attempts within these republics. Only in 1991 did the Baltic republics attempted to secede, and when they were allowed to, the USSR disintegrated.

This can be viewed as another change in the similarity function: in 1989 the experience of satellite-but-independent states such as Poland and Czechoslovakia didn't seem relevant to the Baltics, because the latter were part of the USSR. That is, there was a variable – "being a part of the USSR" – which was apparently deemed relevant even after "post-Glasnost" proved important. Taking these two variables into consideration, the post-Glasnost experience of independent satellite states did not appear to be relevant to the USSR republics. However, when there was a precedent among the Baltics, the variable "being a part of the USSR" dropped out of the similarity function, and the rest of the USSR republics could rely on the same statistics as did the independent states in 1989.

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.

4.1.2 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.

4.1.3 Example: Role Models

Our analysis of precedents, such as President Obama, provides an formal model of the impact of "role models." It has long been argued that students belonging to a minority might rationally decide not to attempt to enter a given professional career, requiring a costly investment in studies, unless they have sufficient evidence that access to that profession is not subject to discrimination. Role models, i.e., minority members with a successful career in that profession, can provide evidence of such lack of discrimination.[14] Our learning model provides an explanation of how the beliefs about chances of success in a profession ($y = 1$) by a member of a minority (i.e., an individual with $x^j = 1$) are formed, how the presence of discrimination is assessed (i.e., in which cases the value of x^j will be considered relevant for prediction), and how precedents of successful professionals with similar features can make

[14] See Chung (2000) and references therein, and Bayer and Rouse (2016).

beliefs more optimistic and hence encourage minority members to attempt entry in a given profession.

4.2 Non-Binary Variables

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 given the case of Obama than it would have been without this case, without necessarily being similar to the latter.[15]

4.3 Similarity Over Variables

Our focus is on similarity between cases, and how it is learnt. But similarity can also be perceived among variables. For example, one might argue that the precedent of President Obama may make it more likely that a woman be elected president. Clearly, a non-white male candidate isn't very similar to a white female one, as far as "race" and "gender" are concerned. Further, even if the variable "race" is no longer perceived as relevant, it doesn't make a non-white man more similar to a white woman than to a white man. However, people might reason along the lines of, "Now that a non-white president was

[15] 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.

elected, why not a woman?" Capturing such reasoning would require generalizing the model described above, allowing a similarity function between variables. For example, "race" and "gender" are similar in that both are in the category of "perceptual variables that were used to discriminate against sub-groups, and that are frowned upon as source of discrimination in modern democracies". Due to this similarity, a change in the weight of one variable, learnt from the empirical similarity as in this paper, may be reflected also in the weight of another variable.

To consider another example, let us revisit the example of the collapse of the USSR (4.1.1 above). One might argue that the variables "Being a part of the Soviet Bloc", "Being a part of the USSR", and "Being a part of Russia" bore some a priori similarity to each other. They seem to be distinct, as the collapse of the Soviet Bloc didn't immediately proceed to the disintegration of the USSR itself. Yet, it is possible that the former inspired the latter, two years later. This might be captured by the variable similarity notion. Moreover, if Chechen rebels felt encouraged by the collapse of the Soviet Bloc *and* of the USSR, they might have been following an inductive process that involved variables before involving cases. Specifically, if, our of the three variables two were proved unimportant, one might be justified in assuming that the third one would follow suit, and make predictions based on a similarity function that does not take it into account.

Observe that the similarity over variables will also be partly learnt from the data. In the latter example, the a priori similarity between the three variables involving the USSR had to be updated given the results of the Chechen uprising. Clearly, such sophisticated forms of learning are beyond the scope of the present paper.

4.4 Statistical Games and Other Classes of Games

Statistical games are reminiscent of "Congestion Games" (Rosenthal, 1973) in that a player's payoff depends only on a summary statistics of the others'

choices. In a classical example, only the frequency of choice of each act matters, rather than the identity of the players choosing it. This is akin to our model, in which only the summary statistic y_i matters for a player's payoff. However, in our case the period game need not be symmetric, and it might be meaningless to consider the frequency of choices of players (or to sum up their chosen variables), as their sets of moves might be unrelated to each other.

Statistical games are similar to Correlated Equilibria (Aumann, 1974) in that we assume that Nature sends a signal to each player before the game is played. However, in our context the signal is commonly known. Thus, any equilibrium of the large game has to induce an equilibrium in each period game (given the realization of x). In this sense our correlating signals, x, bring to mind "Sunspots" (Cass and Shell, 1983). In particular, if one imposes the additional assumption that in a statistical game the x's are payoff-irrelevant, they do function like sunspots, as mere public correlation devices. Viewed thus, our suggestion to use second-order induction to find the similarity function can be considered a theory of sunspot selection.

When considered as a method of equilibrium selection in coordination games, statistical games cannot fail to remind one of "Global Games" (Carlsson and van Damme, 1993). As in the latter, our approach attempts to relate a game to a larger class of games, and to allow the wider context aid in equilibrium selection. However, in Global Games equilibria are chosen ex ante, simultaneously for all games, whereas in statistical games they are chosen sequentially, highlighting the role of statistical learning. Global Games rely on some uncertainty about the game played, while in statistical games, at each period i, G_i is commonly known among its players, and the variables x_i only serve as a coordination device.

4.5 Extensions of Similarity-Nash Equilibria

4.5.1 Iterative Best Response

In some examples one needs more than one step of best-response reasoning to arrive at an equilibrium. For example, consider a modified version of the sequence of revolutions described in Section 3.2. Suppose that $f(\alpha) = \alpha^2$ and that there is a continuum of heterogeneous players where player h's payoff is given by

Payoff to h	$y_i = 1$	$y_i = 0$
$a^h = 1$	$1 + \varepsilon^h$	0
$a^h = 0$	0	$1 - \varepsilon^h$

and $\varepsilon^h \sim U(-1, 1)$, so that her best response is to join the revolution attempt if and only if she thinks that the probability of success is at least $\frac{1-\varepsilon^h}{2} \sim U(0, 1)$. For any initial belief $p_{s_i}(y_i = 1) = p_0 \in (0, 1)$, the best response would be to join the revolution for a fraction $\alpha_0 = p_0$ of the population and not to join it for the remaining fraction. This in turn would generate beliefs $p_1 = p^2 < p_0$, to which the best response would be to join the revolution for an analogous fraction of the population. A formal analysis of such a game would require a generalized notion of Similarity-Nash equilibrium, allowing for an iterative process of best-response to initial beliefs. Such an iterative process would converge to an equilibrium with $\alpha = 0$ for any initial belief $p \in (0, 1)$.

This process brings to mind Level-k reasoning, where one does not start the process with an arbitrary, say, uniform distribution, but with the statistical one obtained from the empirical similarity weighted frequencies.

Note that an iterative process of best responses is at the heart of equilibrium selection in Global Games (Carlsson and van Damme, 1994). Thus, an extension of our equilibrium selection to iterative best responses can simultaneously generalize Global Games (by allowing different games) and our analysis above.

4.5.2 Initial Beliefs

Selecting equilibria by (one shot or iterative) best responses to initial beliefs can be applied to other classes of games. Indeed, one may start with any beliefs $p \in \sum$ about players' strategies, and define a strategy profile $\sigma \in \sum$ to be k-level p-rationalizable if it can be obtained as best-response of degree k to p. If there is some reason to believe that p makes sense as an initial, non-strategic beliefs, a Nash equilibrium $\sigma \in \sum$ that is k-level p-rationalizable may be a more likely prediction than equilibria that aren't (or that can only be obtained by longer chains of reasoning).

Note that such an equilibrium selection procedure would require initial beliefs about the play of the game by each player. By contrast, our definition applies this idea only to statistical games, in which (i) one has to specify initial beliefs only about the variable y_i (and not the entire profile of moves); and (ii) there exists a payoff-irrelevant history that suggests a natural candidate for the initial beliefs.

5 Appendix: Proofs

Whenever needed, we use partial derivatives to derive inequalities. In doing so we obviously extend the definition of the function $\Delta(L, W, l, w)$ to all non-negative real numbers (L, W, l, w) by the function's algebraic formula, whenever well-defined.

Proof of Proposition 1:

Let there be given $w > 1$. We wish to prove that for any $L, W > 2$, $\Delta(L, W, l, 0) < 0$ (where the case $l = 0$, $w > 1$ is obviously symmetric).

The SSE's are given by

$$SSE(\varnothing) = (L + l)\left(-\frac{W}{l + L + W - 1}\right)^2 + W\left(1 - \frac{W - 1}{l + L + W - 1}\right)^2$$

and

$$SSE(\{j\}) = L\left(-\frac{W}{L + W - 1}\right)^2 + W\left(1 - \frac{W - 1}{L + W - 1}\right)^2$$

(where the sub-database for which $x^j = 1$ yields $SSE = 0$). Straightforward calculation yields

$$\Delta(L, W, l, 0) = -Wl\frac{\left(L(W - 2) + (W - 1)^2\right) l + (L + W - 1)(L(W - 2) + W(W - 1))}{(L + W - 1)^2 (l + L + W - 1)^2}$$

which is clearly negative. □□

Proof of Proposition 2:

We need to show that

(i) If $L < W$, $\Delta(L, W, 1, 0) < 0$ and $\Delta(L, W, 0, 1) > 0$;

(ii) If $L > W$, $\Delta(L, W, 1, 0) > 0$ and $\Delta(L, W, 0, 1) < 0$;

(iii) $\Delta(L, L, 1, 0), \Delta(L, L, 0, 1) > 0$.

We first study $\Delta(L, W, 1, 0)$, and show that it is positive for $L \geq W$ and negative for $L < W$. By symmetry, this will also show that $\Delta(L, W, 0, 1)$ is positive for $L \leq W$ and negative for $L > W$, together completing the proof.

The SSE's are given by

$$SSE\left(\varnothing\right) = W\left(1 - \frac{W-1}{L+W}\right)^2 + (L+1)\left(-\frac{W}{L+W}\right)^2$$

and

$$SSE\left(\{j\}\right) = W\left(1 - \frac{W-1}{L+W-1}\right)^2 + L\left(-\frac{W}{L+W-1}\right)^2 + 0.25$$

(where the sub-database for which $x^j = 1$ yields $SSE = \frac{1}{4}$).

It follows that

$$\Delta\left(L, W, 1, 0\right) = \frac{1}{4\left(L+W-1\right)^2\left(L+W\right)^2}\left[\begin{array}{c} L^4 + L^3\left(4W-2\right) + L^2\left(2W^2+2W+1\right) \\ +L\left(-4W^3+6W^2+2W\right) \\ -3W^4+2W^3+5W^2-4W \end{array}\right] \tag{4}$$

Let $a\left(L, W\right)$ denote the the expression in the square brackets in (the RHS of) equation (4), which clearly has the same sign as $\Delta\left(L, W, 1, 0\right)$. First, we observe that

$$a(L, L) = 4L\left(2L^2 + 2L - 1\right) > 0.$$

This establishes Part (iii), and will also be a useful benchmark for Parts (i) and (ii). Indeed, to prove that $a(L, W) > 0$ (and thus that $\Delta\left(L, W, 1, 0\right) > 0$) for $L > W$, we will consider the partial derivative of $a(L, W)$ relative to its first argument, and show that it is positive for $L \geq W$. (Clearly, $a(L, W)$ is a polynomial in its two arguments, and it is well-defined and smooth for all real values of (L, W).) To see this, observe that

$$\begin{aligned} \frac{\partial a\left(L, W\right)}{\partial L} &= 4L^3 + 12L^2W - 6L^2 + 4LW^2 + 4LW + 2L - 4W^3 + 6W^2 + 2W \qquad (5) \\ &= 4L^3 + \left(12W - 6\right)L^2 + \left(4W^2 + 4W + 2\right)L + \left(-4W^3 + 6W^2 + 2W\right) \end{aligned}$$

Observe that $12W - 6 > 0$ (as $W > 2$), and thus the only negative term in this derivative is $-4W^3$. However, for $L \geq W$ it is true that $4LW^2 - 4W^3 \geq 0$ and thus, for $L \geq W$ we have $\frac{\partial a(L,W)}{\partial L} > 0$. Because, for $L \geq W$, $a\left(L, W\right)$

is strictly increasing in L and $a(L,L) > 0$, we also have $a(L,W) > 0$ for $L > W$.

We now turn to the case $L < W$, where equation (5) might be negative (and, indeed, will become negative if L is held fixed and $K \to \infty$.) Again the strategy of the proof is to use direct evaluation at a benchmark and partial derivative arguments beyond, though a few special cases will require attention. The benchmark we use is the case $W = L + 1$. Here direct calculations yield $a(L, L+1) = -4L(2L^2 - 1) < 0$.

This time we consider the partial derivative of $a(L,W)$ wrt to its second argument, and would like to establish that it is negative. If it were, increasing K from $(L+1)$ further up would only result in lower values of $a(L,W)$, and therefore the negativity of $a(L,W)$ (and of $\Delta(L,W,1,0)$) for $L < W$ would be established.

Consider, then,

$$\begin{aligned}
\frac{\partial a(L,W)}{\partial W} &= 4L^3 + 4L^2W + 2L^2 - 12LW^2 + 12LW + 2L - 12W^3 + 6W^2 + 10W - 4 \\
&= 4L^3 + (4W+2)L^2 + \left(12W - 12W^2 + 2\right)L + \left(6W^2 - 12W^3 + 10W - 4\right) \\
&< 4W^3 + (4W+2)W^2 + 12W^2 + 2W - 12LW^2 + 6W^2 - 12W^3 + 10W - 4 \\
&< 4W^3 + (4W+2)W^2 + 12W^2 + 2W + 6W^2 - 12W^3 + 10W - 4 \\
&= -4\left(-3W - 5W^2 + W^3 + 1\right) \qquad (6)
\end{aligned}$$

where the first inequality follows from the fact that $L < W$ and the second from the fact that $L, W > 0$.

We now observe that expression (6) is negative for $W \geq 6$, and thus the partial derivative $\frac{\partial a(L,W)}{\partial W}$ is indeed negative for all $W \geq 6$, $L < W$. Coupled with the fact that $a(L, L+1) < 0$, we obtain $a(L,W) < 0$ for all $W \geq 6$ (and $2 < L < W$).

We now wish to show that $a(L,W) < 0$ holds also for lower values of W. However, as $W > L > 2$ only a few pairs of values (L,W) are possible: $(3,4)$,$(3,5)$,$(4,5)$. Direct calculation shows that $a(L,W)$ is negative for all

these pairs. Specifically, $a(3,4) = -204$, $a(3,5) = -1,424$, and $a(4,5) = -496$.This concludes the proof of Parts (i) and (ii). □□

It will turn out to be convenient to prove Proposition 4 before Proposition 3:

Proof of Proposition 4

It will be convenient to extend the definition of Δ to real-valued arguments, and use calculus. We will only resort to (first- and second- order) partial derivatives with respect to the last two arguments. Note that for positive integers L, W, l, w, the SSE formulae are

$$SSE(\varnothing) = (L+l)\frac{(W+w)^2}{(L+W+l+w-1)^2} + (L+l)^2\frac{W+w}{(L+W+l+w-1)^2}.$$

$$SSE(\{j\}) = LW\frac{L+W}{(L+W-1)^2} + lw\frac{l+w}{(l+w-1)^2}$$

It is therefore natural to define, for positive integers L, W, and any $l, w \in \mathbb{R}$,

$$\begin{aligned}\Delta(L,W,l,w) &= LW\frac{L+W}{(L+W-1)^2} + lw\frac{l+w}{(l+w-1)^2} \\ &\quad - (L+l)\frac{(W+w)^2}{(L+W+l+w-1)^2} - (L+l)^2\frac{W+w}{(L+W+l+w-1)^2}\end{aligned}$$

as long as $l + w \neq 1 - (L+W)$ and $w \neq 1 - l$. Clearly, the function Δ is a rational function in its four arguments, and apart from these points of singularity, it is well-defined and smooth. Note that we are interested in l, w that are positive integers, hence $l, w \geq 1$. In particular, $l + w \geq 2$ while $1 - (L+W) < -3$ and $w \geq 1$ while $1 - l \leq 0$, so that none of the two singular points of Δ is within or even on the boundary of the range of values that is of interest to the statement of the proposition. However, these points will prove useful in analyzing the function.

Next, because our focus is on the behavior of Δ as we change its fourth argument, starting from the critical point $w = \frac{lW}{L}$, it will simplify notation

if we shift the fourth variable to center it around that point. Formally, let $\omega \in R$ and define a function $b : \mathbb{Z}_+^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$b(L, W, l, \omega) = \Delta\left(L, W, l, \frac{lW}{L} + \omega\right).$$

The statements in the Proposition are about the value of the $\Delta(\cdot)$ function evaluated at points where the third argument is a positive integer and the fourth argument is an integer larger or equal than $\left\lceil \frac{lW}{L} \right\rceil$. It is therefore useful to notice that for any positive integers L, W, l, and integer z we can write

$$\Delta\left(L, W, l, \left\lceil \frac{lW}{L} \right\rceil + z\right) = \Delta\left(L, W, l, \frac{lW}{L} + \varepsilon + z\right) = b(L, W, l, z + \varepsilon) \quad (7)$$

where $\varepsilon = \left\lceil \frac{lW}{L} \right\rceil - \frac{lW}{L}$. Note that $\varepsilon \in [-0.5, 0]$ if $\left\lceil \frac{lW}{L} \right\rceil = \left\lfloor \frac{lW}{L} \right\rfloor$ and $\varepsilon \in [0, 0.5)$ if $\left\lceil \frac{lW}{L} \right\rceil = \left\lceil \frac{lW}{L} \right\rceil$.

Since the Proposition assumes $w = \left\lceil \frac{lW}{L} \right\rceil \geq 1$, it has to be the case that $\frac{lW}{L} > 0.5$ and $-\frac{lW}{L} < -0.5$.

We prove the proposition as follows:

(1) We first show that $b(L, W, l, \omega)$ is strictly decreasing in ω for $\omega \geq 1$ (Lemma 1);

(2) Next, we prove that $b(L, W, l, \omega)$ has a limit as $\omega \to \infty$ and that it is a negative number (Lemma 2);

(3) Direct calculation shows that $b(L, W, l, 1.5) > 0$, and from this we conclude that, as a function of ω, $b(L, W, l, \omega)$ has a unique root larger than 1.5 (Lemma 3);

(4) We prove that $b(L, W, l, \omega) > 0$ for $\omega \in [-0.5, 1.5]$ (Lemma 4);

(5) Next, we show that $\frac{\partial b(L,W,l,\omega)}{\partial l} > 0$ for $\omega \geq 2$ (Lemma 5);

(6) We then show that, for all $l' > l > 1$, $\widetilde{w} > \frac{l'W}{L}$, if $\Delta(L, W, l, \widetilde{w}) \geq 0$ then $\Delta(L, W, l', \widetilde{w}) \geq 0$ (Lemma 6).

Before we proceed to formally state and prove these lemmas, let us explain why they prove the result:

Part (i) follows from (4): we need to show that (for all $L, W > 2, l, w > 0$), we have $\Delta(L, W, l, w), \Delta(L, W, l, w + 1) > 0$. In terms of the function b,

$\Delta(L,W,l,w) = b(L,W,l,\varepsilon)$ and $\Delta(L,W,l,w+1) = b(L,W,l,\varepsilon+1)$. Thus we have to show that $b(L,W,l,\varepsilon), b(L,W,l,\varepsilon+1) > 0$ where $\varepsilon = \left[\frac{lW}{L}\right] - \frac{lW}{L} \in [-0.5, 0.5)$. Clearly, this follows from Lemma 4.

Part (ii) follows from (1) and (3) because b is a smooth function of ω in the range $\omega \geq 1$.

Part (iii) follows from (6): If l' is such that $\left[\frac{l'W}{L}\right] \geq w^*(L,W,l) - 2$, the claim follows from the fact that $w^*(L,W,l') \geq \left[\frac{l'W}{L}\right] + 2$. Thus we focus on the case $\left[\frac{l'W}{L}\right] < w^*(L,W,l) - 2$**.**

Using part (i) and the definition of w^*, $\Delta(L,W,l,q) \geq 0$ for any integer q such that $0 \leq q \leq w^*(L,W,l) - 1$. Claim (6) implies that for the same values of q, $\Delta(L,W,l',q) \geq 0$. It follows that the smallest integer w'' ($w'' > \left[\frac{l'W}{L}\right]$) for which $\Delta(L,W,l',w'')$ becomes negative is greater or equal than $w^*(L,W,l)$ and thus $w^*(L,W,l') \geq w^*(L,W,l)$.

Finally, to see Part (iv), assume that W/L is an integer, and consider integers $l' > l > 1$. Let $w = \left[\frac{lW}{L}\right]$ and $w' = \left[\frac{l'W}{L}\right]$, that is, $w = \frac{lW}{L}$ and $w' = \frac{l'W}{L}$ as these are integers. Then, Lemma 5 implies that, if $b(L,W,l,\omega) = \Delta(L,W,l,w+\omega) > 0$ for $\omega \geq 2$, then $b(L,W,l',\omega) = \Delta(L,W,l',w'+\omega) > 0$ (for the same ω). It follows that the smallest integer ω ($\omega > 1$) for which $\Delta(L,W,l',w'+\omega)$ becomes negative is bigger than that for which $\Delta(L,W,l,w+\omega)$ becomes negative and thus $w^*(L,W,l') - w' \geq w^*(L,W,l) - w$.

We start by providing the explicit formula for $b(L,W,l,\omega)$:

$$b(L,W,l,\omega) = \frac{LW(L+W)}{(L+W-1)^2} + \frac{l(lW+L\omega)[l(L+W)+L\omega]}{[lW+L(l+\omega-1)]^2} \qquad (8)$$
$$-\frac{(l+L)(lW+LW+L\omega)(lL+L^2+lW+LW+L\omega)}{(-L+lL+L^2+lW+LW+L\omega)^2}$$

This is a rational function in ω, with two vertical asymptotes where either the denominator of the first term or the denominator of the third term in 8

vanishes. We denote these singular points by $\underline{\omega}$ and $\overline{\omega}$, respectively:

$$\begin{aligned}\overline{\omega} &= 1-\frac{l\,(L+W)}{L}=1-l-\frac{lW}{L}<0\\ \underline{\omega} &= 1-\frac{(l+L)\,(L+W)}{L}<\overline{\omega}\end{aligned}$$

Thus, for $\omega>\overline{\omega}$, $b\,(L,W,l,\omega)$ is a smooth function.

We can now establish:

Lemma 1 $b\,(L,W,l,\omega)$ is strictly decreasing in ω for $\omega\geq 1$.

Proof:

Differentiate $b\,(L,W,l,\omega)$ with respect to ω:

$$\begin{aligned}\frac{\partial b\,(L,W,l,\omega)}{\partial\omega} &= \frac{(2L(l+L)(lW+L(W+\omega))(l(L+W)+L(L+W+\omega)))}{(L^2+lW+L(-1+l+W+\omega))^3}\\ &\quad -\frac{(L(l+L)(l(L+2W)+L(L+2(W+\omega))))}{(L^2+lW+L(-1+l+W+\omega))^2}\\ &\quad +\frac{(lL^2(-2lW+l^2(L+W)+lL(-1+\omega)-2L\omega))}{(lW+L(-1+l+\omega))^3}\end{aligned}$$

The above expression can be rewritten as

$$-\frac{L^3\left[z_0(L,W,l)+z_1(L,W,l)\omega+z_2(L,W,l)\omega^2+z_3(L,W,l)\omega^3+z_4(L,W,l)\omega^4\right]}{(lW+L(l+\omega-1))^3(L^2+lW+L(l+W+\omega-1))^3} \tag{9}$$

where $z_0(L,W,l), z_1(L,W,l), z_2(L,W,l), z_3(L,W,l), z_4(L,W,l)$ are defined as:

$$\begin{aligned}z_0(L,W,l) &= -2l^4(L-W)(L+W)^3-l^2L^2(L+W)^2(6+L(2L-9)-2W^2)\\ &\quad -2l^3L(L+W)^2(L(2L-3)-2W^2)\\ &\quad +lL^3\left[L(2+3(L-2)L)+4W+6(L-2)LW+3(+L-2)W^2\right]\\ &\quad +L^4\left[2W-L(L+W-1)\right]\end{aligned}$$

$$z_1(L,W,l)=L\left\{\begin{array}{c} L^3\left[(2(l-1)^4+4(l-1)^3L+(3-4l+2l^2)L^2\right]\\ +W\left[\begin{array}{c}6(l-1)l(2-l+l^2)L^2\\ +6(2l-1)(1-l+l^2)L^3\\ +3(1-2l+2l^2)L^4+6lL(l+L)(1+l^2+lL)W\\ +2l(l+L)(2l+l^2+L+lL)W^2\end{array}\right]\end{array}\right\}$$

$$z_2(L,W,l) = 3L^2\left\{2l^3W^2 + L\left[\begin{array}{c}(-2+4l-4l^2+2l^3)L + L^2\left[2-4l+3l^2+(l-1)L\right]\\ +[(4l(1-l+l^2)+2L+l(6l-4)L+(2l-1)L^2]W\\ +(3l^2+lL)W^2\end{array}\right]\right\}$$

$$z_3(L,W,l) = L^3\left[L^3 + 2l(3l-2)W + L^2(-4+6l+W) + L(6-8l+6l^2-2W+6lW)\right]$$

$$z_4(L,W,l) = L^4(-2+2l+L)$$

First, notice that L^3 and the denominator of expression (9) are strictly positive, hence the sign of (9) is equal to the opposite sign of the polynomial in ω on its numerator. Second, notice that $z_1(L,W,l)$, $z_2(L,W,l)$, $z_3(L,W,l)$, and $z_4(L,W,l)$ are strictly positive for all admissible values of $\{L,W,l\}$. It follows that the derivative of the polynomial in ω on the numerator of (9) is strictly positive for positive values of ω. Hence, if we can show that the polynomial is positive for some positive value of ω, then it is positive for all larger values of ω as well. Finally, we evaluate the polynomial at $\omega = 1$ and show that it is positive.

$$\begin{aligned} & z_0(L,W,l) + z_1(L,W,l)\,(1) + z_2(L,W,l)\,(1) + z_3(L,W,l)\,(1) + z_4(L,W,l)\,(1) \\ = \;& 2l(l+L)(L+W)^3[L^2 + l^2W + lL(2+W)] > 0 \end{aligned}$$

This allows us to conclude that $\frac{\partial b(L,W,l,\omega)}{\partial \omega} < 0$ for all $\omega \geq 1$. □□

Lemma 2 $\exists \lim_{\omega\to\infty} b\,(L,W,l,\omega) < 0$.

Proof:

Observe that

$$\begin{aligned} \lim_{\omega\to\infty} b\,(L,W,l,\omega) &= \frac{LW(L+W)}{(L+W-1)^2} + l - l - L \\ &= \frac{-L(L-1)^2 - (L-2)LW}{(L+W-1)^2} < 0. \end{aligned}$$

Which concludes the proof of the lemma. □

Lemma 3 $b\,(L,W,l,\omega)$ has exactly one root in $\omega \in (1.5,\infty)$.

Proof:

We know that the singular points of b are negative. This means that for $\omega \geq 0$, $b(L, W, l, \omega)$ is a smooth function. Further, algebraic calculations[16] show that $b(L, W, l, 1.5) > 0$ for all $L, W > 2$, $l > 1$ such that $\left[\frac{lW}{L}\right] \geq 1$. $b(L, W, l, 1.5) > 0$ and we established that $b(L, W, l, \omega) < 0$ for ω large enough. Hence it has to have a root at some $\omega > 1.5$. Further, it is unique because b is strictly decreasing in ω over this range. □

Lemma 4 $b(L, W, l, \omega) > 0$ for $\omega \in [-0.5, 1.5]$.

Proof:

We need to consider two cases.

Case 1: $l = 1$

In this case, the vertical asymptotes are at $\underline{w} = -\frac{W}{L} - (W + L)$ and $\overline{\omega} = -\frac{W}{L} < -0.5$ (the inequality holds because it must be true that $\left[\frac{lW}{L}\right] = \left[\frac{W}{L}\right] \geq 1$) so for $\omega \geq -0.5$ the function is smooth. Algebraic calculations[17] show that for $l = 1$, for all $L, W > 2$ such that $\left[\frac{lW}{L}\right] \geq 1$, $\frac{\partial b(L,W,l,\omega)}{\partial \omega}$ is strictly negative for all $\omega \geq -0.5$. This, together with the fact that $b(L, W, l, 1.5) > 0$, proves that $b(L, W, l, \omega) > 0$ for $\omega \in [-0.5, 1, 5]$.

Case 2: $l > 1$

Algebraic calculations[18] show that for $l > 1$, for all $L, W > 2$ such that $\left[\frac{lW}{L}\right] \geq 1$, $b(L, W, l, -0.5) > 0$. To study the sign of $b(L, W, l, \omega)$ for $\omega \in [-0.5, 1.5]$ we observe that it is positive at $\omega = -0.5$ and at $\omega = 1.5$, and that it is continuous on the interval. Thus, to prove that it is positive throughout the interval it suffices to show that it has no roots in it.

Observe that $b(L, W, l, \omega)$ is a rational function in ω with a fourth degree polynomial (in ω) in its numerator. Every root of b is a root of this polynomial, and thus b can have at most four real roots. We claim that it has at least one real root in each of the following intervals:

[16] Available upon request. (Part (a) in the Appendix for referees)
[17] Available upon request. (Part (c) in the Appendix for referees)
[18] Available upon request. (Part (b) in the Appendix for referees).

(a) $(\underline{\omega}, \overline{\omega})$

(b) $(\overline{\omega}, -0.5)$

(c) $(1.5, \infty)$.

To see that there is a root in (a), observe that

$$\lim_{\omega \to_+ \overline{\omega}} b(L, W, l, \omega) = \lim_{\omega \to_- \overline{\omega}} b(L, W, l, \omega)$$

$$= \frac{LW(L+W)}{(L+W-1)^2} - \frac{L^2 l(l-1)}{0} - \frac{L(L+l)(L+LW-Ll)(L+W+1)}{L^2(L+W)^2} = -\infty$$

$$\lim_{\omega \to_+ \underline{\omega}} b(L, W, l, \omega)$$

$$= \frac{LW(L+W)}{(L+W-1)^2} + \frac{l[-L(L+W+l-1)][-L(L+W-1)]}{L^2(L+W)^2}$$

$$- \frac{-L^2[l(L+2l-1)+l(l-1)]}{0^+} = +\infty$$

Thus, b, which is continuous over $(\underline{\omega}, \overline{\omega})$, goes from $+\infty$ to $-\infty$ and has to cross 0 over the interval.

As for (b), observe, again, that $\lim_{\omega \to_+ \overline{\omega}} b(L, W, l, \omega) = -\infty$ and that $b(L, W, l, 0.5) > 0$.

Finally, (c) has been established in Lemma 3.

We can now consider the interval of interest, $[-0.5, 1.5]$. We know that b is positive at the two endpoints. If it were non-positive at some point over this interval, the numerator of b would have to have two roots in the interval – either two distinct roots or a multiple one. In either case, we would have a total of five real roots for a polynomial of degree 4, which is impossible, and thus we conclude that b is strictly positive throughout $[-0.5, 1.5]$. □

Lemma 5 $b(L, W, l, \omega)$ is strictly increasing in l for $\omega \geq 2$.

Proof:

The derivative of $b(L, W, l, \omega)$ wrt l is:

$$L^3 \frac{\zeta_0(L, W, \omega) + \zeta_1(L, W, \omega) l + \zeta_2(L, W, \omega) l^2 + \zeta_3(L, W, \omega) l^3}{(-L + lL + lW + L\omega)^3(-L + lL + L^2 + lW + LW + L\omega)^3} \qquad (10)$$

where $\zeta_0(L,W,\omega)$, $\zeta_1(L,W,\omega)$, $\zeta_2(L,W,\omega)$, $\zeta_3(L,W,\omega)$ are defined as:

$$\zeta_0(L,W,\omega) = L^3(\omega-1)\left(\begin{array}{c} L^3\omega^2 + W(4(\omega-1)^2\omega + W^2(2\omega-1) + 3W(1-3\omega+2\omega^2)) \\ +L^2(3(\omega-1)\omega^2 + W(2\omega(1+\omega)-1)) \\ +L\left(\begin{array}{c} 2(\omega-1)^2\omega(1+\omega) + W^2(\omega(4+\omega)-2) \\ +3W(1+\omega(-3+\omega+\omega^2)) \end{array}\right) \end{array}\right)$$

$$\zeta_1(L,W,\omega) = L^2\left(\begin{array}{c} W^2(12W(\omega-1)^2 + W^2(2\omega-3) + 6(\omega-1)^2(2\omega-1)) \\ +L^4(\omega-2)\omega + 3L^2(2(\omega-1)^2\omega^2 + 4W(\omega-1)^2(1+\omega) + W^2(\omega^2-3)) \\ +LW(-6+6W(\omega-1)^2(4+\omega) + 6\omega(4-4\omega+\omega^3) + W^2(-9+\omega(4+\omega))) \\ +L^3(6(\omega-1)^2\omega + W(-3+\omega(3\omega-4))) \end{array}\right)$$

$$\zeta_2(L,W,\omega) = 3L(L+W)^2\left(\begin{array}{c} L\left(L(\omega-2)+2(\omega-1)^2\right)\omega \\ +W^2(2\omega-3) + W\left(4(\omega-1)^2 + L(\omega^2-3)\right) \end{array}\right)$$

$$\zeta_3(L,W,\omega) = 2(L+W)^3(L(\omega-2)\omega + W(2\omega-3))$$

First, notice that L^3 and the denominator of expression (10) are strictly positive. Second, notice that $\zeta_0(L,W,\omega)$, $\zeta_1(L,W,\omega)$, $\zeta_2(L,W,\omega)$, $\zeta_3(L,W,\omega)$ are strictly positive for all admissible values of $\{L,W\}$ and $\omega \geq 2$. Since l is an integer. it follows that the polynomial in l on the numerator of (10) is strictly positive for $\omega \geq 2$. This allows us to conclude that $\frac{\partial b(L,W,l,\omega)}{\partial \omega} > 0$ for all $\omega \geq 2$. □

Lemma 6 For all $l' > l > 1$, $\widetilde{w} > \frac{l'W}{L}$, if $\Delta(L,W,l,\widetilde{w}) \geq 0$ then $\Delta(L,W,l',\widetilde{w}) \geq 0$.

Proof:

If $\widetilde{w} = \left[\frac{l'W}{L}\right]$ or $\widetilde{w} = \left[\frac{l'W}{L}\right] + 1$, the conclusion $\Delta(L,W,l',\widetilde{w}) \geq 0$ follows from Part (i).

Assume, then, that $\widetilde{w} \geq \left[\frac{l'W}{L}\right] + 2 \geq \left[\frac{lW}{L}\right] + 2$. Recall that $w = \left[\frac{lW}{L}\right]$ with $\varepsilon = \left[\frac{lW}{L}\right] - \frac{lW}{L}$ and denote $w' = \left[\frac{l'W}{L}\right], \varepsilon' = \left[\frac{l'W}{L}\right] - \frac{l'W}{L}$. Next, let $\omega = \widetilde{w} - w$ and $\omega' = \widetilde{w} - w'$. Thus

$$\widetilde{w} = w + \omega = \frac{lW}{L} + \varepsilon + \omega = w' + \omega' = \frac{l'W}{L} + \varepsilon' + \omega'$$

Clearly, as $l' > l$, we have $w' \geq w$ and therefore $\varepsilon' + \omega' \leq \varepsilon + \omega$. Note that $\omega, \omega' \geq 2$ and thus $\omega + \varepsilon, \omega' + \varepsilon' \geq 1$.

We assume that

$$\Delta\left(L,W,l,\widetilde{w}\right)=\Delta\left(L,W,l,w+\omega\right)=b\left(L,W,l,\omega+\varepsilon\right)\geq 0$$

and need to show

$$\Delta\left(L,W,l',\widetilde{w}\right)=\Delta\left(L,W,l',w'+\omega'\right)=b\left(L,W,l',\omega'+\varepsilon'\right)\geq 0.$$

Indeed, $b\left(L,W,l,\omega+\varepsilon\right)\geq 0$, coupled with Lemma 5, implies that $b\left(L,W,l',\omega+\varepsilon\right)\geq 0$. Further, as $\omega'+\varepsilon'\leq\omega+\varepsilon$, Lemma 1 (with $\omega+\varepsilon,\omega'+\varepsilon'\geq 1$) implies that $b\left(L,W,l',\omega'+\varepsilon'\right)\geq 0$, which completes the proof of the lemma. □□

Proof of Proposition 3

The proof relies on the analysis used to prove Proposition 4. Let us denote by $\bar{l}$ the closest integer to $\frac{L}{W}$ $(=\frac{wL}{W}$ because we deal with the case $w=1)$, that is, $\bar{l}=\left[\frac{L}{W}\right]$.

We need to show that, for every $0<l\leq\left\lfloor\frac{L}{W}\right\rfloor+1$, $\Delta\left(L,W,l,1\right)>0$. Recalling the symmetry of Δ with respect to values of y, $\Delta\left(L,W,l,1\right)=\Delta\left(W,L,1,l\right)$. Thus, we need to show that $\Delta\left(W,L,1,l\right)>0$ for all $0<l\leq\left\lfloor\frac{L}{W}\right\rfloor+1$.

In (7) we had

$$\Delta\left(L,W,l,\left[\frac{lW}{L}\right]+z\right)=\Delta\left(L,W,l,\frac{lW}{L}+\varepsilon+z\right)=b\left(L,W,l,z+\varepsilon\right)$$

which, replacing L and W, as well as l and w, yields

$$\Delta\left(W,L,w,\left[\frac{wL}{W}\right]+z\right)=\Delta\left(W,L,w,\frac{wL}{W}+\varepsilon+z\right)=b\left(W,L,w,z+\varepsilon\right)$$

and by setting $w=1$, also

$$\Delta\left(W,L,1,\left[\frac{L}{W}\right]+z\right)=\Delta\left(W,L,1,\frac{L}{W}+\varepsilon+z\right)=b\left(W,L,1,z+\varepsilon\right)$$

For $0<l\leq\left\lfloor\frac{L}{W}\right\rfloor+1$, denoting $z=l-\bar{l}$ we have $l=\bar{l}+z=\left[\frac{L}{W}\right]+z$. We can then write

$$\Delta\left(W,L,1,l\right)=\Delta\left(W,L,1,\left[\frac{L}{W}\right]+z\right)=\Delta\left(W,L,1,\frac{L}{W}+\varepsilon+z\right)=b\left(W,L,1,z+\varepsilon\right)$$

where $\varepsilon = \left\lceil \frac{L}{W} \right\rceil - \frac{L}{W} \in [-0.5, 0.5)$ and $z \in \{1 - \left\lceil \frac{L}{W} \right\rceil, ..., 1\}$ if $\left\lceil \frac{L}{W} \right\rceil = \left\lfloor \frac{L}{W} \right\rfloor$ and $z \in \{1 - \left\lceil \frac{L}{W} \right\rceil, ..., 0\}$ if $\left\lceil \frac{L}{W} \right\rceil = \left\lfloor \frac{L}{W} \right\rfloor + 1$.

Denoting the fourth argument of b by $\omega = z + \varepsilon$, we observe that, because $z \geq 1 - \left\lceil \frac{L}{W} \right\rceil$, $\omega \geq 1 - \frac{L}{W}$. Further, as $z \leq 1$ and $\varepsilon < 0.5$, $\omega < 1.5$. Thus, it suffices to show that $b(W, L, 1, \omega) > 0$ for $\omega \in \left[-\frac{L}{W} + 1, 1.5\right]$. However, we know that $b(W, L, 1, \omega)$ is continuous and differentiable for $\omega > -\frac{L}{W}$, that $\frac{\partial b(W,L,1,\omega)}{\partial \omega} < 0$ for all $\omega \geq -\frac{L}{W}$, and that $b(W, L, 1, 1.5) > 0$. Therefore, $b(W, L, 1, \omega) > 0$ for all $\omega \geq -\frac{L}{W}$. This concludes the proof. □

6 References

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

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

Argenziano, R. and I. Gilboa (2017), "Learning What is Similar: Precedents and Equilibrium Selection", *working paper*.

Aumann, R. (1974), "Subjectivity and Correlation in Randomized Strategies", *Journal of Mathematical Economics*, **1**: 67–96.

Bayer, A. and Rouse, C. E. (2016) "Diversity in the Economics Profession: A New Attack on an Old Problem", *Journal of Economic Perspectives*, **30:** 221–242.

Carlsson, H. and Van Damme, E. (1993), "Global Games and Equilibrium Selection", *Econometrica*, **61**: 989-1018.

Cass, D. and K. Shell (1983), "Do Sunspots Matter?", *Journal of Political Economy*, **91**: 193–228.

Chung, K. S. (2000) "Role Models and Arguments for Affirmative Action", *American Economic Review*, **90**: 640-648.

Edmond, C. (2013). Information manipulation, coordination, and regime change. Review of Economic Studies, 80(4), 1422-1458.

Fryer, R. and M. O. Jackson (2008), "A Categorical Model of Cognition and Biased Decision Making", *The B.E. Journal of Theoretical Economics*, 8.

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

Halaburda, H., Jullien, B., & Yehezkel, Y. (2016). Dynamic competition with network externalities. *working paper*

Harsanyi, J. C. and R. Selten (1988), *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.

Medin. D. L. and M. M. Schaffer (1978), "Context Theory of Classification Learning", *Psychological Review*, **85**: 207-238.

Nagel, R. (1995), "Unraveling in Guessing Games: An Experimental Study", *American Economic Review*, **85**: 1313–1326.

Nosofsky, R. M. (1984), "Choice, Similarity, and the Context Theory of Classification", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**: 104-114.

Nosofsky, R. M. (1988), "Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**: 700-708.

Nosofsky, R. M. (2011), "The Generalized Context Model: An Exemplar Model of Classification", in *Formal Approaches in Categorization*, Cambridge University Press, New York, Chapter 2, 18-39.

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

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

Rosenthal, R. W. (1973), "A Class of Games Possessing Pure-Strategy Nash Equilibria", *International Journal of Game Theory*, **2**: 65–67.

Schelling, Th. C. (1960), *The Strategy of Conflict*. Cambridge: Harvard University Press

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

Selten, R. (1977), "The Chain Store Paradox", *Theory and Decision*, **9**: 127-159.

Shepard, R. N. (1957), "Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space", *Psychometrika*, **22**: 325-345

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

Stahl, D. O. and P. W. Wilson (1995), "On Players' Models of Other Players: Theory and Experimental Evidence", *Games and Economic Behavior*, **10**: 213-254.

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

7 Appendix for Referees

a) Calculation of $b\left(L,W,l,1.5\right)>0$.

We evaluate the function $b\left(L,W,l,\omega\right)$ at $\omega=1.5$ and find the following expression:

$$\frac{L*g\left(L,W,l\right)}{(-1+L+W)^2(L+2lL+2lW)^2(L+2lL+2L^2+2lW+2LW)^2}$$

where $g\left(L,W,l\right)$ can be expressed as a polynomial in W :

$$\begin{aligned}
& g\left(L,W,l\right) \\
= \; & (32l^4+64l^3L+32l^2L^2)W^5 \\
& +16l\left[2L^3+l^3(8L-1)+2l^2L(1+8L)+lL^2(5+8L)\right]W^4 \\
& +4lL\left[12l^3(4L-1)+3lL^2(21+16L)+L^2(4+27L)+8l^2(-1+3L+12L^2)\right]W^3 \\
& +2L^2\left[\begin{array}{c} -3(L-2)L^2+8l^4(8L-3)+16l^3(-2+3L+8L^2)+ \\ 2l^2(-6-6L+69L^2+32L^3)+2lL(6-L+33L^2) \end{array}\right]W^2 \\
& +\left[\begin{array}{c} 16l^4(2L-1)+3L(2+5L-4L^2)+32l^3(-1+L+2L^2) \\ +4l^2(-3-12L+29L^2+8L^3)+4l(-6+15L-14L^2+17L^3) \end{array}\right]L^3W \\
& -3L^4\left(L-1\right)^2\left(3+2L\right)+12l^2L^4\left(L-1\right)^2+12lL^4\left(L-1\right)^3
\end{aligned}$$

Notice that for $W,L>2$ and $l>0$ the terms multiplying W^5, W^4, and W^3 are positive. The terms multiplying W^2 and L^3W and the constant are polynomials in l. For $l>0$, all three are increasing in l, as the coefficients of the positive powers of l are positive. Moreover, all three are positive when evaluated are $l=1$, hence for all $l>1$ as well. In particular, the coefficient of W^2 evaluated at $l=1$ is equal to $-68+112L+270L^2+127L^3>0$. The coefficient of L^3W evaluated at $l=1$ is equal to $-84+82L+139L^2+88L^3>0$. Finally, the constant evaluated at $l=1$ is equal to $3L^4\left(2L-3\right)\left(L-1\right)^2>0$.

We have proved that $g\left(L,W,l\right)>0$. Since $\frac{L}{(-1+L+W)^2(L+2lL+2lW)^2(L+2lL+2L^2+2lW+2LW)^2}>0$, this concludes the proof.

b) Calculation of $b\left(L,W,l,-0.5\right)>0$ for $l>1$.

We evaluate the function $b\left(L,W,l,w\right)$ at $w=-0.5$ and find the following expression:

$$\frac{-L*h\left(L,W,l\right)}{(-1+L+W)^2(-3L+2lL+2lW)^2(-3L+2lL+2L^2+2lW+2LW)^2}$$

where $h\left(L,W,l\right)$ can be expressed as a polynomial in L :

$$\begin{aligned}
& h\left(L,W,l\right) \\
= \; & (20l-18)\,L^7+\left[-4l^2(8W-5)+l(-76+92W)+45-36W\right]L^6 \\
& +\left[-36+4(23-10l)l+(81-4l(90+l(-75+16l)))W-2(9-78l+64l^2)W^2\right]L^5 \\
& +\left[\begin{array}{c} 9-36l+20l^2+(-54+388l-528l^2+224l^3-32l^4)W \\ +(36-492l+780l^2-256l^3)W^2+(116l-192l^2)W^3 \end{array}\right]L^4 \\
& +\left[\begin{array}{c} -4l(18-43l+24l^2-4l^3)-4l(-74+234l-168l^2+32l^3)W \\ -4l(52-185l+96l^2)W^2-4l(32l-8)W^3 \end{array}\right]WL^3 \\
& -8l^2W^2\left[-19+56W-30W^2+4W^3+l^2(-6+24W)+l(24-84W+32W^2)\right]L^2 \\
& -16l^3W^3\left[6-3l+(8l-14)W+4W^2\right]L-16l^4W^4(2W-1)
\end{aligned}$$

In what follows, we prove that $h\left(L,W,l\right)<0$ for all $l>0$ and $L,W>2$. The constant term is negative. The coefficient of L is negative because it is the product of a negative term and a quadratic expression in W with a positive coefficient on the square which is positive and increasing at $W=2$, hence for any larger W too. Similarly, the coefficient of L^2 is negative because it is the product of a negative term and a quadratic expression in l with a positive coefficient on the square which is positive and increasing at $l=2$, hence for any larger l too.

The coefficient of L^3 is the product of W, which is positive, and a third degree polynomial in W which can be shown to be negative in the relevant range. In particular, the polynomial has a negative coefficient on the third and second power. At $W=2$, this polynomial is equal to $-56l+236l^2-288l^3-240l^4$ which is negative for all $l>1$. Moreover, its derivative at $W=2$ is equal to $-152l+488l^2-864l^3-128l^4$ which is also negative for all

$l > 1$. Finally, the fact that this derivative is negative $W = 2$ implies that it is also negative for all values of $W > 2$, because the negative coefficients on the third and second powers of W guarantee that the function is concave in W for positive W.

The coefficient of L^4 is a third degree polynomial in W which can be shown to be negative in the relevant range ($l > 1$, $W > 2$). The polynomial has a negative coefficient on the third power. Evaluated at $W = 2$, it takes value $45-300l+548l^2-576l^3-64l^4 < 0$ for all $l > 1$. Moreover, its derivative wrt W evaluated at $W = 2$ is equal to $90 - 188l + 288l^2 - 800l^3 - 32l^4$ which is also negative for all $l > 1$. Finally, its second derivative wrt W is equal to $-8\left(-9 + 123l - 195l^2 + 64l^3 + (144l - 87)lW\right)$ which is negative at $W = 2$ and decreasing in W for all positive values of W.

The coefficient of L^5 is a quadratic function of W with a negative coefficient on the square, which is negative and decreasing at $W = 3$, hence negative for all larger values of W too. The coefficient of L^6 is a quadratic function of l with a negative coefficient on the square, which is positive for $l = 2$ and negative for all larger values of l. The coefficient of L^7 is positive.

Since the coefficient L^7 is positive, and we want to prove that the whole polynomial in L is negative, we prove that the sum of the terms in L^7 and L^5 is negative.

First, notice that the condition $\frac{lW}{L} \geq \frac{1}{2}$ implies that $L \leq 2lW$, which in turn implies:

$$(20l - 18)\, L^7 < 4\,(20l - 18)\, L^5 l^2 W^2$$

which in turn implies that

$$\begin{aligned}
& (20l - 18)\, L^7 + \left[\begin{array}{c} -36 + 4(23 - 10l)l \\ +(81 - 4l(90 + l(-75 + 16l)))W - 2(9 - 78l + 64l^2)W^2 \end{array}\right] L^5 \\
< \;& 4\,(20l - 18)\, L^5 l^2 W^2 + \left[\begin{array}{c} -36 + 4(23 - 10l)l \\ +(81 - 4l(90 + l(-75 + 16l)))W - 2(9 - 78l + 64l^2)W^2 \end{array}\right] L^5 \\
= \;& \left[\begin{array}{c} (80l - 72)\, l^2 W^2 - 36 + 4(23 - 10l)l \\ +(81 - 4l(90 + l(-75 + 16l)))W - 2(9 - 78l + 64l^2)W^2 \end{array}\right] L^5 \\
= \;& \left[\left(92l - 40l^2 - 36\right) + \left(300l^2 - 64l^3 - 360l + 81\right) W + (-128l^2 + 236l - 90)W^2\right] L^5
\end{aligned}$$

The last expression is a quadratic in W which is negative for all $W > 2$. In particular, it has a negative coefficient on the square, hence it is concave. Evaluated at $W = 2$ it is equal to $-128l^3+48l^2+316l-234 < 0$ for all $l > 1$. Moreover, its derivative evaluated at $W = 2$ is equal to $-64l^3-212l^2+584l-279 < 0$ for all $l > 1$.

To conclude the proof that the whole polynomial in L is negative, we still need to address the fact that the coefficient of L^6 is positive at $l = 2$. In particular, we do so by proving that the sum of the terms in L^6 and L^4 is negative at $l = 2$. First, notice that the condition $\frac{lW}{L} \geq \frac{1}{2}$ implies that $L \leq 2lW$, which in turn implies:

$$\begin{aligned} & \left[-4l^2(8W-5)+l(-76+92W)+45-36W\right]/_{l=2}L^6 \\ < \;& 4\left[-4l^2(8W-5)+l(-76+92W)+45-36W\right]/_{l=2}L^4l^2W^2 \end{aligned}$$

which in turn implies that

$$\begin{aligned} = \;& \left[-4l^2(8W-5)+l(-76+92W)+45-36W\right]/_{l=2}L^6 \\ & +L^4\left[\begin{array}{c} 9-36l+20l^2+(-54+388l-528l^2+224l^3-32l^4)W \\ +(36-492l+780l^2-256l^3)W^2+(116l-192l^2)W^3 \end{array}\right]/_{l=2} \\ < \;& 4\left[-4l^2(8W-5)+l(-76+92W)+45-36W\right]/_{l=2}L^4l^2W^2 \\ & +L^4\left[\begin{array}{c} 9-36l+20l^2+(-54+388l-528l^2+224l^3-32l^4)W \\ +(36-492l+780l^2-256l^3)W^2+(116l-192l^2)W^3 \end{array}\right]/_{l=2} \\ = \;& \left(-216W^3-308W^2-110W+17\right)L^4 < 0 \text{ for all } W > 2. \end{aligned}$$

This concludes the proof that $b\left(L,W,l,-0.5\right) > 0$ for $l > 1$.

c) Calculation of $\frac{\partial b(L,W,l,w)}{\partial w} < 0$ for all $w \geq -\frac{W}{L}$ for the case $l = 1$

For $l = 1$, the $b\left(L,W,l,w\right)$ function and its derivative with respect to w are

$$\begin{aligned} b\left(L,W,1,w\right) = & \frac{LW(L+W)}{(L+W-1)^2}+\frac{L+W+Lw}{W+Lw} \\ & -\frac{(1+L)(W+LW+Lw)(L+L^2+W+LW+Lw)}{(L^2+W+LW+Lw)^2} \end{aligned}$$

$$\frac{\partial b\left(L,W,1,w\right)}{\partial w}=\frac{-L^{3}\phi\left(L,W,w\right)}{\left(W+Lw\right)^{2}\left(L^{2}+W+LW+Lw\right)^{3}}$$

where $\phi\left(L,W,w\right)$ is the following cubic expression in w in which all the coefficients, including the constant, are positive.

$$\begin{aligned}&\phi\left(L,W,w\right)\\ =\;& L^{5}+3L^{3}W+3L^{4}W+4LW^{2}+8L^{2}W^{2}+4L^{3}W^{2}+2W^{3}+4LW^{3}+2L^{2}W^{3}\\ &+w\left(3L^{4}+8L^{2}W+10L^{3}W+2L^{4}W+4LW^{2}+6L^{2}W^{2}+2L^{3}W^{2}\right)\\ &+w^{2}\left(4L^{3}+2L^{4}+L^{5}+2L^{2}W+3L^{3}W+L^{4}W\right)+w^{3}L^{4}\end{aligned}$$

The sign of the coefficients guarantees that the expression is positive, for all $w\geq 0$. To examine the sign of $\phi\left(L,W,w\right)$ for $w\in[-\frac{W}{L},0)$, notice that:

a) $\phi\left(L,W,-\frac{W}{L}\right)=L^{2}(L+W)^{3}>0$

b) $\phi\left(L,W,0\right)=L^{5}+3L^{3}W+3L^{4}W+4LW^{2}+8L^{2}W^{2}+4L^{3}W^{2}+2W^{3}+4LW^{3}+2L^{2}W^{3}>0$

c)

$$\begin{aligned}\frac{\partial\phi\left(L,W,w\right)}{\partial w} &= \left(3L^{4}+8L^{2}W+10L^{3}W+2L^{4}W+4LW^{2}+6L^{2}W^{2}+2L^{3}W^{2}\right)\\ &\quad+2w\left(4L^{3}+2L^{4}+L^{5}+2L^{2}W+3L^{3}W+L^{4}W\right)+3L^{4}w^{2}\\ &> \left(3L^{4}+8L^{2}W+10L^{3}W+2L^{4}W+4LW^{2}+6L^{2}W^{2}+2L^{3}W^{2}\right)\\ &\quad+2w\left(4L^{3}+2L^{4}+L^{5}+2L^{2}W+3L^{3}W+L^{4}W\right)\\ &> \left(3L^{4}+8L^{2}W+10L^{3}W+2L^{4}W+4LW^{2}+6L^{2}W^{2}+2L^{3}W^{2}\right)\\ &\quad-2\frac{W}{L}\left(4L^{3}+2L^{4}+L^{5}+2L^{2}W+3L^{3}W+L^{4}W\right)\\ &= 3L^{3}\left(L+2W\right)>0\end{aligned}$$

where the first inequality follows from the fact that $3L^{4}w^{2}>0$ and the second from the fact that $w>-\frac{W}{L}$.

Hence we can conclude that $\phi\left(L,W,w\right)$ is positive and increasing in the whole interval $\left(-\frac{W}{L},0\right)$, hence the function $b\left(L,W,1,w\right)$ is decreasing for all $w>-\frac{W}{L}$.

Second-Order Induction: Uniqueness and Complexity*

Rossella Argenziano† and Itzhak Gilboa‡

August 2018

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 "empirical similarity" 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 best similarity function 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.

Keywords: Empirical Similarity, Belief Formation.

*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. r.argenziano@essex.ac.uk

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

1 Introduction

Where do beliefs come from? How do, and how should economic agents estimate the likelihood of future events? Decision theory remains mostly silent on this point. The axiomatic foundations laid by Ramsey (1926a,b), de Finetti (1931,1937), Savage (1954), and Anscombe-Aumann (1963) are very powerful in arguing that rational individuals should behave as if they had probabilistic beliefs (to be used for expected utility maximization), and arguably also that actual economic agents behave this way. But they shed no light over the question of the selection of prior probabilities. In a sense, they deal with form but not with content.

The natural answer to the belief formation problem is provided by equilibrium analysis: whether in games or in markets, rational agents' beliefs are assumed to coincide with the modeler's. However, equilibria need not be unique. And, more fundamentally, one needs to ask whether agents' behavior will converge to an equilibrium in the first place, which brings us back to the belief formation question. In short, it appears that there is a need for theories of belief formation that would be (i) relatively general and applicable to a variety of economic settings; (ii) sufficiently rational to credibly apply to weighty economic decisions; and (iii) intuitive enough to be thought of as idealized models of the way actual people think.

In the quest for reasonable models of belief formation, two fellow disciplines might be of help: statistics and psychology. The former has a normative flavor, while the latter – descriptive. Statistics and, more recently, machine learning attempt to develop effective ways of prediction based on past data, with no claim to describe the way people think. By contrast, psychology aims at modeling human reasoning, be it more or less rational. Recent developments in cognitive science highlight a promising bridge between these disciplines: a specific class of learning techniques developed in statistics and machine learning, namely kernel methods and support vector machines, are closely related to 'exemplar learning' models developed in psychology: "...kernel methods have neural and psychological plausibility, and theoretical results concerning their behavior are therefore potentially relevant for human category learning." (Jaekel, Schoelkopf, and Wichmann, 2009, p. 381). This paper presents a model of belief formation based both on kernel techniques and on insights from the exemplar learning literature.

We start by assuming that the probability of a future event is estimated by its

similarity-weighted relative frequency in the past.[1] More explicitly, given past observations $(x_i, y_i)_{i \leq n}$ (where x_i is a vector of real-valued predictors and y_i is the indicator of the event in question), and a new point x_p, the probability of the event occurring next is estimated by

$$P(y_p = 1) = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)} \tag{1}$$

where s is a non-negative similarity function defined on pairs of x vectors. When all past events are deemed equally relevant, probability is estimated by empirical frequency. But in general past occurrences are weighted by their similarity: more similar circumstances gain higher weight than less similar ones. This estimation is referred to as *first-order induction*.

The second level of learning involves finding a similarity function, to be used in (1), from the data as well. Specifically, we consider a Leave-One-Out cross-validation technique: each similarity function is assessed by the sum of squared errors it would have yielded, were it to be used in sample, to predict each y_i based on the other observations. A function that brings this sum of squared errors to a minimum is referred to as an "empirical similarity", and it is used here as an obviously idealized model of the way people learn which features are more important than others to assess similarities. Because this process deals with learning how first-order induction should be performed, it will be dubbed *second-order induction*.

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 arguments of the similarity function. 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 similarity functions 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. Rational agents might therefore end up using different, suboptimal similarity functions.

There are many modeling choices to be made, in terms of the nature of the vari-

[1] We follow the convention in psychology and decision theory to label kernel functions as 'similarity functions.'

ables (the predictors and the predicted), as well as of the similarity function. We study here two extreme cases: in the "binary" model all the variables take only the values $\{0,1\}$, and so does the similarity function. Further, we consider only similarity functions that are defined by weights in $\{0,1\}$: each variable is either taken into account or not, and two observations are similar (to degree 1) if and only if they are equal on all the relevant variables. In the "continuous" model, by contrast, all variables (predictors and predicted) are continuous, and the similarity function is allowed to take any non-negative value as well. We focus on functions that are exponential in weighted Euclidean distances where the weights are allowed to be non-negative extended real numbers.

In both models we find the same qualitative conclusions: (i) 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 empirical similarity will learn the functional relationship. The similarity function is likely to be unique, but even if it is not, different empirical similarity functions would provide the same predictions (Propositions 2 and 4). By contrast (ii) If the number of predictors is large relative to the number of observations, it is highly probable that the empirical similarity will not be unique (Propositions 3 and 5). Further, (iii) If the number of predictors is not bounded, the problem of finding the empirical similarity is NPC (Theorems 1 and 2).

To see the implications of these results, let us contrast two prediction problems: in the first, an agent tries to estimate the probability of his car being stolen. In the second, the probability of success of a revolution attempt. In the first problem, there are several relevant variables to take into account, such as the car's worth, the neighborhood in which it is parked, and so forth. One can think of the number of these variables as relatively limited. By contrast, the number of observations of cars that were or were not stolen is very large. In this type of problems it stands to reason that empirical similarity be unique. Further, as the number of variables isn't large, the complexity result has little bite. Thus, different people are likely to come up with the same similarity functions, 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 may be very large. Researchers may come up with novel perspectives on

a given history, and suggest new military, economic, and sociological variables that might help judge which historical cases are similar to which. In this type of examples our results suggest that the optimal similarity function may not be unique, and that, even if it is unique, people may fail to find it. That is, to the extent that second-order induction describe a psychological process people implicitly go through, they may learn to judge similarity by functions that are not necessarily the best one. It follows that they may also not find the same function (even if the "best" one is unique). 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.

The rest of this paper is organized as follows. The next subsection discusses first- and second-order induction, and the specific formulas we use, in the literatures in statistics, psychology, and decision theory. Section 2 deals with the questions of monotonicity, uniqueness, and computational complexity of the empirical similarity function in the binary model, while Section 3 provides the counterpart analysis for the continuous model. Section 4 concludes with a general discussion.

1.1 Related Literature

Using similarity-weighted averages is an intuitive idea that appeared in statistics as "kernel methods" (Akaike, 1954, Rosenblatt, 1956, Parzen, 1962). Further, it has also been suggested that the "best" kernel function be estimated from the data. In particular, Nadaraya (1964) and Watson (1964) suggested to find the optimal bandwidth of the kernel (see also Park and Marron, 1990). Our focus is mostly on the qualitative question, namely, which variables to include in the function, rather than on the quantitative one, that is, how close is "close". The question of optimal bandwidth is obviously of interest in applied statistical work, but for the purposes of economic modeling we find the choice of variables to be of greater import. Be that as it may, we are unaware of results about optimal kernel functions that are along the lines of our results here.

Cortes and Vapnik (1995) suggested the widely-used method of "support vector machines" (SVMs) for classification problems. This technique is based on the idea that if a simple linear classifier might not exist in the original space, there might still be one in a higher dimensional space. The latter is often taken to be the kernel functions defined by points in the learning database, resulting in kernel classification

coupled with optimization of the coefficients of the kernel function, and of the functions itself. This technique is also used to estimate probabilities (see Vapnik, 2000) along lines that are similar to logistic regression. We are unaware of results in this literature that are similar to ours.

The formula (1) also appeared in the psychological literature, in the Generalized Context Model (Medin and Schaffer, 1978, and Nosofsky, 1984). In this domain the task that participants in an experiment are asked to perform is typically a classification task (to guess whether $y_p = 1$ or $y_p = 0$), rather than probability assessment (that is, to provide a number in $[0, 1]$ for the probability that $y_p = 1$). However, when modeling the frequency with which participants classify a new case as $y_p = 1$ or $y_p = 0$, it appears that these frequencies are given by (1). In particular, the model finds that classification of a new 'exemplar' is based on the similarity between the latter and a set of training exemplars, with a mental process that resembles our notion of first order induction. Exemplars are represented as points in a multidimensional psychological space and the similarity between any two is a decreasing function of their distance in this space (the Multidimensional Scaling Approach, see Shepard, 1957,1987). Importantly, Nosofsky (1988) finds that people seem to learn the relative importance of different attributes in the similarity function in a process that resembles what we call second order induction. (See Nosofsky, 2014, for a survey). The fact that, for classification problems, the same formula appeared in machine learning and in psychology was noted by Jaekel, Schoelkopf, and Wichmann (2008, 2009). Yet, formal analysis of optimal similarity functions, whether for classification or for probability estimation, seems to be lacking.

Similarity-based classification was axiomatized in Gilboa and Schmeidler (2003), and similarity-weighted probability estimation as in (1) was axiomatized in Billot, Gilboa, Samet, and Schmeidler (2005) and in Gilboa, Lieberman, and Schmeidler, [GLS] (2006) (the former for the case of y being a discrete variable with at least 3 values, the latter for the case of two values discussed here). GLS (2006) 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.[2] Lieberman (2010, 2012) analyzed the asymptotic properties of

[2]The learning process presented here has been suggested and analyzed in GLS (2006) as a statistical technique. However, in this paper our focus is descriptive, and we use the model to describe human reasoning. In this sense our paper is similar to Bray (1982), who considers a statistical technique, namely OLS, as a model of economic agents' reasoning.

such estimators. (See also Lieberman and Phillips, 2014, 2017). The asymptotic results in this literature assume a given DGP (typically, using a formula such as (1), with a noise variable, as the "true" statistical model), whereas our results are more agnostic about the underlying DGP.

In sum, both the formula (1) and the notion of learning the optimal similarity function to be used within it, have appeared in psychology, statistics and machine learning, and decision theory. Given the independent derivation of the same idea in first two disciplines, which are very different in terms of their goals, these notions of first- and second-order induction hold a promise for modeling beliefs of economic agents. The statistical pedigree suggests that this mode of belief formation is not irrational in any obvious and systematic way; the psychological ancestry indicates that it is not too far from what human beings might conceive of.

2 A Binary Model

2.1 Case-Based Beliefs

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[3]. We focus here on prediction by rather basic case-based formulae.[4] 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.

We assume that prediction is made based on a similarity function $s : X \times X \to$

[3]Luckily, the two questions are not divorced from each other. For example, linear regression has 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 "exemplar learning" in psychology, and various kernel techniques ended up being identical to similarity-based techniques axiomatized in decision theory. (See Gilboa and Schmeidler, 2012.)

[4]As in Gilboa and Schmeidler (2001, 2012).

$\mathbb{R}_+$. 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 one. The similarity function should incorporate not only intrinsic similarity judgments, but also judgments of relevance, probability of recall and so forth.[5]

In this section we present 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.

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 the similarity weighted average[6]

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)} \tag{2}$$

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). Because the similarity function 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.

[5] 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.

[6] Gilboa, Lieberman, and Schmeidler (2006) provide axioms on likelihood judgments (conditioned on databases) that are equivalent to the existence of a function s such that (6) holds for any database B. Billot, Gilboa, Samet, and Schmeidler (2005) consider the similarity-weighted averaging of probability vectors with more than two entries.

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

$$s_w(x_i, x_p) = \prod_{\{j|w^j=1\}} \mathbf{1}_{\left\{x_i^j = x_p^j\right\}} \tag{3}$$

(where $s_w(x_i, x_p) = 1$ for all (x_i, x_p) if $w^j = 0$ for all j.) 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 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 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 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 mean squared error it would have had.

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 mean squared error to be[7]

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

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 $MSE(J)$ 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[8]. Formally, we will refer to the set of indices of these cases, 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

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

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

$$\overline{y}^{(J,z)} = \frac{\sum_{i \in b(J,z)} y_i}{|b(J, z)|} \tag{4}$$

[7] Similar results would hold for other loss functions. See subsection 4.1.

[8] Splitting the database into such bins is clearly an artifact of the binary model. We analyze a more realistic continuous model in Section 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

$$\overline{y}_i^{(J,z)} = \frac{\sum_{k \in b(J,z), k \neq i} y_k}{|b(J,z)| - 1} \tag{5}$$

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

Given the predictions $\overline{y}_i^{(J,z)}$, we can now calculate $MSE(J)$ for all the possible similarity functions. We will not, however, stop here and select the similarity function that minimizes the mean squared error as the "empirical similarity". There is one more element to consider. In choosing a subset of variables to be included in J, 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.[9] There are two types of considerations leading to such a preference. 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[10]. Moreover, one may argue that such preference for a smaller set of predictors is evolutionarily selected partly due to the statistical normative considerations. 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 would

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

[10] As a normative theory, the preference for simple theories is famously attributed to William of Ockham (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). As a descriptive theory, the preference for simplicity appears in Wittgenstein's Tractatus (1922) at the latest.

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.*

We now turn to analyze the properties of the empirical similarity, to address the question of whether we should expect rational agents with access to a common database to agree on their predictions.

2.3 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: MSE can be non-monotone with respect to set inclusion.[11] 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 $m = 1$. Consider the following database and the corresponding MSE's of the subsets of the variables:

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

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 $y_i = 0$ (which is compared to $\frac{a}{a+b-1} > \frac{a}{a+b}$). In both cases, the 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.

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

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* $i = k (\mathrm{mod}\, n)$.

Consider 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, it is still positive. 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 smaller 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* $J \subseteq M \setminus \{j\}$ *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. Formally,

Definition 3 *A variable* $j \in M$ is a function of $J \subseteq M \setminus \{j\}$ 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,

Proposition 1 Assume that j is informative relative to $J \subseteq M \setminus \{j\}$ 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, $MSE(J \cup \{j\}) < MSE(J)$. Conversely, if j is not informative relative to J, then for any t-replica of B, $MSE(J \cup \{j\}) \geq MSE(J)$, 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* $J \subseteq J' \subseteq M \setminus \{j\}$. *It is possible that* j *is informative for* J, *but not for* J' *as well as vice versa.*

2.4 Uniqueness

We have seen in section 2.3 that monotonicity of the MSE is not generally guaranteed. Immediate implications are that the best fit is not necessarily achieved by a unique subset of variables J, and in particular by the full set of all available predictors ($J = M$). For concreteness, consider the following database

Example 2 Let $n = 12$ and $m = 2$. Consider the following database and the corresponding MSE's of the subsets of the variables:

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

J	$MSE(J)$
$\varnothing$	0.2975
$\{1\}$	0.2
$\{2\}$	0.2
$\{1,2\}$	0.3333

Thus, the set of variables that minimize the MSE and the $AMSE$ need not be unique.[12] Observe that the different similarity functions will also differ in their predictions, both in-sample and certainly also out-of-sample. To see that, let us begin with the prediction for observations $i = 1, 2$. In these, $x_i^1 = 1$ and $x_i^2 = 0$. The similarity function s_J that corresponds to $J = \{1\}$ yields an estimated y value of $\overline{y}_i^{s_J} = 0.8$ whereas the similarity $s_{J'}$ for $J' = \{2\}$ yields $\overline{y}_i^{s_{J'}} = 0.2$. Thus, even though the two similarity functions obtain the same MSE, and this is the minimal one over all such functions, their predictions for 4 out of the 12 observations *in the sample* are very different. Clearly, two such functions can also disagree over the predictions out of sample. In fact, they can disagree on out of sample observations even if they fully agree in sample, for example, if in the sample two variables have the same informational content. Specifically, if in the sample $x^1 = x^2$, for any $c > 0$ the optimal similarity function will not include both variables. Assume that it includes one of them. Then there are at least two similarity functions that minimize the MSE and that are indistinguishable over all the observations in the sample. Yet, if a new observations would have $x_p^1 \neq x_p^2$, these two functions might well disagree.

This raises the issue of when can we reasonably expect rational agents faced with the same prediction problem to adopt the same empirical similarity. In this section, we derive two results that characterize sufficient conditions for the two possible cases. Proposition 2 identifies a class of prediction problems for which including all available predictors in the similarity function does indeed minimize the MSE, hence the $AMSE$ too as long as the cost c is sufficiently small. At the other extreme, Proposi-

[12] In this example we only compute the MSE, and the minimizers are the two singletons. Clearly, for a small enough c these two subsets are also the minimizers of the $AMSE$.

tion 3 identifies a class of prediction problems for which at least two disjoint subsets of variables minimize MSE and $AMSE$. The comparison between the conditions in Propositions 2 and 3 sheds light on features of a prediction problem that make agreement among rational agents more or less likely to occur.

Let us first consider data generating processes that are conducive to the inclusion of all variables in the empirical similarity. Assume that the values of the predictors, (x_i) are i.i.d. with a joint distribution g on X, and that $y_i = f(x_i)$ for a fixed $f : X \to \{0,1\}$.[13] Let us refer to this data generating process as (g, f). We introduce the following definition, then present our result:

Definition 4 *A variable $j \in M$ is* informative for (g, f) *if there are values $z^{-j} = \left(z^k\right)_{k \neq j}$ such that (i) $f(z^{-j} \cdot 0) \neq f(z^{-j} \cdot 1)$; and (ii) $g(z^{-j} \cdot 0), g(z^{-j} \cdot 1) > 0$, where $z \cdot q \in X$ is the vector obtained by augmenting z^{-j} with $z^j = q$ for $q \in \{0,1\}$.*

Proposition 2 Assume a data generating process (g, f) where all $j \in M$ are informative for (g, f). Then there exists $\bar{c} > 0$ such that, for all $c \in (0, \bar{c})$,

$$P\left(\arg\min_{J \subseteq M} AMSE(J, c) = \{M\}\right) \to_{n \to \infty} 1$$

The proposition thus says that, if there is an underlying relationship so that the distribution of y_i is a function of x_i, but x_i alone, and this function remains constant for all observations, then, with a large enough database (i.e. fixing m and allowing n to grow) the only set of predictors that minimize the $AMSE$ is the set of all predictors – unless some of them are not informative.

By contrast, let us now consider the other extreme case, where n is fixed and m is allowed to grow. In this case, under fairly general probabilistic assumptions, we find the opposite conclusion, namely that non-uniqueness is the rule rather than the exception. Formally, fix n and (letting m grow) assume that for each new variable j, and for every $i \leq n$,

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

[13] One may consider more general cases in which y is random, and $P(y_i = 1 \mid x_i) = f(x_i)$ for some $f : X \to [0,1]$. In this case one can prove results that are similar to Proposition 4 below.

for a fixed $\eta \in (0, 0.5)$. 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". For such a process we can state:

Proposition 3 For every $n \geq 4$, every $c \geq 0$, and every $\eta \in (0, 0.5)$, if there are at least two cases with $y_i = 1$ and at least two with $y_i = 0$, then

$$P\left(\begin{array}{c} \exists J, J' \in \arg\min_{J \subseteq M} AMSE\left(J, c\right), \\ J \cap J' = \varnothing \end{array}\right) \rightarrow_{m \rightarrow \infty} 1$$

Proposition 2 can be viewed as dealing with a classical scientific problem, where the set of relevant variables is limited, and many observations are available, perhaps even by active experimentation. In this case we would expect that all informative variables would be used in the optimal similarity function (if the fixed cost per variable is sufficiently low). Thus the set of optimal functions will be a singleton, defined by the set M, and, in particular, different people who study the same database are likely to converge on the same similarity function and therefore on the same predictions for any new data point x_p. By contrast, Proposition 3 deals with cases that are more challenging to scientific study: the number of observations is fixed – which suggests that active experimentation is ruled out – and also considered to be small relative to the number of predictors that may be deemed relevant. The data generating process in Proposition 3 can be viewed as a model of a process in which people come up with additional possible predictors for a given set of cases. For example, 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 are considered "best", and, as a result, different people may entertain different beliefs about future observations based on the same data.

2.5 Complexity

Examples in which different sets of variables obtain precisely the same, minimal $AMSE$ might be knife-edge, hence disagreement might appear to be unlikely to occur in practice. In this section, we present a second reason why rational agents faced with the same prediction problem might adopt different similarity functions and disagree in their predictions. As the number of possible predictors in a database grows, so does the complexity of finding the optimal set of variables, even if it is unique. Formally, we 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 1 *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.[14] Clearly, the practical import of this complexity result depends crucially on the number of variables, m.[15] 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.[16]

[14] 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.

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

[16] If we restrict EMPIRICAL-SIMILARITY to accept problems with a bounded m, say, $m \leq m_0$, then it obviously becomes polynomial (in n, involving coefficients of the order of magnitude of 2^m).

3 A Continuous Model

One can 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.,

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)} \tag{6}$$

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).[17] 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 $y_0 \in Y$.[18]

For many purposes it makes sense to consider more general similarity 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

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

where n is a norm on $\mathbb{R}^m$. Indeed, this functional form is often used in explaining psychological data about classification problems.[19] Gilboa, Lieberman, and Schmeidler (2006) and Gayer, Gilboa, Lieberman (2007) also study the case of a weighted

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

$$\hat{y}_p^s \in \arg\max_y \sum_{i \leq n} s(x_i, x_p) \mathbf{1}_{\{y = y_i\}} \tag{7}$$

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

[18] 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.

[19] Shepard (1987) suggests that a similarity function which is exponential in distance (in a "psychological space") might be a 'universal law of generalization.' See Nosofsky (2014) for a more recent survey. Note, however, that the similarity function in that literature is mostly for a classification task, rather than for probability estimation.

Euclidean distance, where

$$s^w(x, x') = \exp\left(-\sum_{j=1}^{m} w^j \left(x^j - x'^j\right)^2\right) \tag{8}$$

with $w_j \geq 0$.[20]

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 in (8) 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 (8) where $w^j = 1$ in (3).) 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 (8) where the summation is taken over all j's such that $w^j < \infty$.

The definition of the empirical similarity extends to this case almost verbatim: the MSE is defined in the same way, and one can consider similarity functions given by (8) for some non-negative $(w^j)_{j \leq m}$. Rather than thinking of $MSE(s)$ as a function of a set of predictors, $J \subseteq M$, denoted $MSE(J)$ as above, one would consider it as a function of a vector of weights, $w = (w^j)_{j \leq m}$, denoted $MSE(w)$. We will similarly define the adjusted MSE 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.

[20] 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. See Lieberman (2010, 2012), and Lieberman and Phillips (2014, 2017). In most of these results the exponential function is assumed, though some results hold more generally.

However, when we think of an empirical similarity as a function s^w that minimizes the $AMSE$, we should bear in mind the following.

Observation 2 *There are databases for which*

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

(This Observation is proved in the Appendix.) The reason that the argmin of the MSE 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 MSE approximately.

We can define approximately optimal similarity: for $\varepsilon > 0$ let

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

Thus, the ε-arg min $AMSE$ is the set of weight vectors that are ε-optimal. We are interested in the shape of this set for small $\varepsilon > 0$.

3.1 Almost-Uniqueness

We argue that the main messages of our results in the binary case carry over to this model as well. Again, the key questions are the relative sizes of n and m, and the potential causal relationships between observations: when there are $n >> m$ independent observations that obey a functional rule $y = f(x)$ – which, in particular, implies that x_i contains enough information to predict y_i – the optimal weights will be unique, and different people are likely to converge to the same opinion. By contrast, when $m >> n$, it is likely that different sets of variable will explain the same (relatively small) set of observations.

Let us first consider the counterpart of (g, f) processes, where the observations (x_i, y_i) are i.i.d. For simplicity, assume 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 $g(z) \geq \eta > 0$ 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[21]

$$y_i = f(x_i).$$

Refer to this data generating process as (g, f).

Proposition 4 Assume a data generating process (g, f) (where f is continuous). Let there be given $\nu, \xi > 0$. There are an integer N_0 and $W_0 \geq 0$ such that for every $n \geq N_0$, the vector w_0 defined by $w_0^j = W_0$ satisfies

$$P(MSE(w_0) < \nu) \geq 1 - \xi.$$

The proposition says that, if there is an underlying relationship so that y_i is a continuous function of x_i, and this function remains constant for all observations, then, when the database is large enough, with very high probability, this relationship can be uncovered. This is a variation on known results about convergence of kernel estimation techniques (see Nadaraya, 1964, Watson, 1964) and it is stated and proved here only for the sake of completeness.[22]

We take Proposition 4 as suggesting that, under the assumption of the (g, f) process, different individuals are likely to converge to similar beliefs about the value of y_p for a new case given by x_p within the known range. The exact similarity function that different people may choose may not always be identical. For example, if $x_i^1 = x_i^2$ for every observation in the database, one function s^w may obtain a near-perfect fit with $w^1 >> 0$ and $w^2 = 0$ and another, $s^{w'}$, – with $w'^1 = 0$ and $w'^2 >> 0$. If one individual uses s^w to make predictions, and another – $s^{w'}$, they will agree on the predicted values for all x that are similar to those they have encountered in the database. In a sense, they may agree on the conclusion but not on the reasoning. But, as long as they observe cases in which $x^1 = x^2$, they will not have major disagreements about any particular prediction.

However, we also have a counterpart of Proposition 3: 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

[21] Similar conclusion would follow if we allow y_i to be distributed around $f(x_i)$ with an i.i.d. error term.

[22] We are unaware of a statement of a result that directly implies this one, though there are many results about optimal bandwidth that are similar in spirit.

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 $\left(x_i^j\right)_{k<i}$, 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, as in Proposition 3, we allow for a rather general class of data generating processes, where, in particular, the x's are not constrained to be independent.[23] The message of the following result is that the empirical similarity is non-unique.

For such a process we can state:

Proposition 5 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 such that for every $n \geq N$ there exists $M(n)$ such that for every $m \geq M(n)$,

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

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 minimal value. However, one could expect this set to be convex, as would be the case if we were considering the minimization of a convex function. This convexity would also suggest a simple follow-the-gradient algorithm to find a global minimum of the $AMSE$ function. But the Proposition states that this is not the case. For $\varepsilon = 0$ we could expect ε-$\arg\min AMSE$ to be a singleton (hence a convex set), but as soon as $\varepsilon > 0$ we will find that there are ε-minimizers of the $AMSE$ whose convex combinations need not be ε-minimizers. Clearly, this is possible because our result is asymptotic: given ε we let n, and then $m \geq M(n)$ go to infinity. But we find the present order of quantifiers to be natural: ε indicates a degree of tolerance to suboptimality, and it can be viewed as a psychological feature of the agent, as can the cost c. The pair (ε, c) can be considered as determining the agent's preferences for the accuracy and simplicity trade-off. An agent with given preferences is confronted with a database, and we ask whether her "best" explanation of the database be unique as more data accumulate. Proposition 5 suggests that multiplicity of local optima of the similarity function is the rule when the number of variables is allowed to increase relative to that of the observations.

[23] 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.

3.2 Complexity

Importantly, our complexity result extends to the continuous 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 2 *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 $w^j = 0$.[24][25]

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.

4 Discussion

4.1 Robustness of the Results

There are a number of modeling decisions to be made in order to state formal results as those above, including the ranges of the variables, of the similarity functions, of the weights therein, as well as the loss functions used to measure the in-sample fit, and the cross validation criterion. Our choices were guided by what seemed the simplest and/or most commonly used definitions, and yet one may wonder how robust are the results.

[24] 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.

[25] 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).

Let us first comment on the ranges of the variables: we study here two extreme cases, one in which all variables are in $\{0,1\}$, and the other in which they are continuous. The former seems best suited to clarify conceptual issues, but it may be oversimplified in some ways. (In particular, in our model similarity is a binary relation which is also transitive.) The latter model is obviously more flexible, but requires messier statements of the results. As the same conceptual results hold in both, one may speculate that this will be the case for various intermediate cases (say, continuous variables with a binary similarity function, or vice versa).

The selection criteria for the optimal similarity function are not crucial for most of our results. In fact, the results are all based on perfect fits: Propositions 2 and 4 state that, with high probability, a perfect fit will be obtained only by including all informative variables, thus resulting in a unique set of variables (in the binary model), or an almost-unique collection of weights (in the continuous one). By contrast, Propositions 3 and 5, which state that, with very high probability the (ε-)optimal similarity function will *not* be unique also rely on perfect fits, only this time a perfect fit that is obtained by disjoint sets of variables. Finally, the complexity results are also based on a perfect fit which is equivalent to a perfect set cover. When perfect fit is involved, most selection criteria agree. In particular, we need a loss function and a cross-validation technique that yields 0 loss if, and *only* if, a perfect fit is obtained in-sample.

The only important assumption for the complexity results (Theorems 1 and 2) is the discontinuity of the $AMSE$ near zero weights. That is, we assume, in a way that's similar to the adjusted R^2 in linear regression, that there is a minimal fixed cost to be paid for the inclusion of a variable (that is, to have a positive weight for that variable). This discontinuity at 0 adds the combinatorial aspect to the $AMSE$ minimization problem, and allows the reduction of combinatorial problems as in our proofs. Our complexity results do not directly generalize to an objective function that is continuous at zero. Furthermore, it is possible that they do not hold in this case.[26] However, as explained above, we find the discontinuous cost function rather reasonable: the difference between a weight $w^j > 0$ and $w^j = 0$ involves the need to collect and recall data about the variable, to use another variable in making computations, and so forth. It seems that some cost is incurred by the inclusion of a

[26] Eilat (2007) proves, in the context of linear regression, that Aragones et al. (2005) complexity result holds if the cost function is discontinuous at zero, but not otherwise.

variable, and that this cost isn't entirely negligible if we think of the model as trying to capture a cognitive process people undergo in trying to make predictions.

4.2 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. One can distinguish among 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.

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 Vapnik, 1998) 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 results are in line with the general message of statistical learning theory. Specifically, our positive learning results, namely, Propositions 2 and 4, assume that there is an underlying functional relationship of the type $y = f(x)$, keep m fixed and let n grow to infinity. The fact that learning is possible under these circumstances may not seem like a major surprise. Observe, however, that our results do not deal

with learning the function f directly and, for that reason, they do not directly follow from results about classes of functions with a low VC dimension. In particular, in our model the prediction of y is always done non-parametrically, by weighted averages of other y values, rather than by some direct function of the x variables. In this context, our learning results should be interpreted as saying that if, unbeknownst to the agent, y is a function of x, but the agent adheres to case-based prediction as she usually does, she is likely to make correct predictions *even though* she is ignorant of the nature of the underlying process.

Our negative results (Propositions 3 and 5) may also 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 results discuss the inevitability of *large* differences in opinion. Finally, our complexity results (Theorems 1 and 2), which also point at inability of learning, seem to have no obvious counterpart in the literature. Importantly, these results show that learning might be difficult even in the simple process discussed here (and justified by psychological research).

4.3 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 Savage, 1954, section 5.5). 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 method of generating a prior belief over the state space.

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.

In this paper we focus on probabilistic beliefs, or point estimates of the variable y given the x's. In case of uniqueness of the similarity function, or at least agreement among all the empirical similarity functions, one may consider these estimates to be objective, and proceed to assume that all rational agents would share them. But in case of disagreement, one may ask whether it is rational for the agents to disagree. For example, if there are multiple similarity functions that obtain a best fit, is it rational for an agent to choose one and based her predictions on that function alone? Wouldn't it more rational for her, assuming unbounded computational ability, to find all optimal functions and somehow take them into account in her predictions? These are valid questions which are beyond the scope of this paper.

4.4 Agreement

Economic theory tends to assume that, given the same information, rational agents would entertain the same beliefs: differences in beliefs can only arise from asymmetric information. 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 Aumann (1976) in the celebrated "agreeing to disagree" result.

The Common Prior Assumption has been the subject of heated debates (see Morris, 1995, Gul, 1998, as well as Brandenburger and Dekel, 1987 in the context of Aumann, 1987). 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 Propositions 2 and 4) can identify economic set-ups where beliefs are likely to be in agreement. By contrast, negative results (such as Propositions 3 and 5) point to problems where agreement is less likely to be the case.

In Argenziano and Gilboa (2018) we apply this approach to equilibrium selection in coordination games. We study in detail the extreme case of adding a single variable to the similarity function in the binary model: assuming that there is agreement about the other set of relevant variables, J, will a new variable $j \notin J$ be added to it? This is about as small as a small world can be, and we interpret our analysis in that paper as shared by all players in the game. By contrast, when the number of variables grows, players may play off-equilibrium due to the negative results proved above.

4.5 Higher-Order Induction Processes

Second-order processes raise questions about yet higher order processes of the same nature, and the possibility of infinite regress. The question then arises, why do we focus on second-order induction and do not climb up the hierarchy of higher-order inductive processes? Higher order induction can indeed be defined in the context of our model. Our notion of second-order induction consists of learning the similarity function from the database of observation. One may well ask, could this learning process be improved upon? For example, we have been using a leave-one-out technique. But the literature suggests also other methods, such as k-fold cross-validation, in which approximately $1/k$ of the database is taken out each time, and their y values

are estimated by the remaining observations. One can consider, for a given database, the choice of an optimal k, or compare these methods to bootstrap methods (see, for instance, Kohavi, 1995). Similarly, kernel methods can be compared to nearest-neighbor methods (Fix and Hodges, 1951, 1952). In short, the process we assume in this paper, of second-order induction, can itself be learnt by what might be called third-order induction, and an infinite regress can be imagined. Isn't restricting attention to second-order induction somewhat arbitrary? Is it a result of bounded rationality?

A few comments are in order. First, in some types of applications lower orders may provide good approximations. For example, suppose that it is indeed the case that $y = f(x)$ as in Propositions 2 and 4. Zero-order induction may refer to the assumption that there is nothing to be learnt from the past about the future, or, at least, that the x variables contain no relevant information. This would surely lead to poor predictions as compared to the learnable process ($y = f(x)$). First-order induction would be using a fixed similarity function to predict y based on its past values. This would provide much better estimates, though also systematic biases (in particular, near the boundaries of the domain of x). Thus, second-order induction is needed, which, in particular, leads to higher weights, and "tighter" similarity functions for large n. This is basically the message of Propositions 2 and 4: similarly to decreasing the bandwidth of the Nadaraya-Watson estimator when n increases, computing the empirical similarity leads (with very high probability) to convergence of the estimator to $y_p = f(x_p)$. Third-order induction could improve these results, say, by making the rate of convergence faster. But it is not needed for the conceptual message of Propositions 2 and 4, and, importantly, of Propositions 3 and 5: for a small m and increasing n we can expect learning to occur, and agreement to result, whereas neither is guaranteed when m is large relative to n. Thus, the marginal contribution of higher orders of induction, in terms of the conceptual import of our results, seems limited.

Second, our model can also be applied to strategic set-ups, such as equilibrium selection in coordination games. In these set-ups the data generating process is partly, or mostly about the reasoning of other agents, and being even one level behind the others may have a big effect of the accuracy of one's predictions, as well as on one's payoff. However, in such a game any reasoning method can be an equilibrium in the "meta-game", in which players select a reasoning method and then use it for predicting others' behavior. For example, players might adopt zero-order induction,

assume that the past is completely irrelevant and make random selections at each period. Thus, zero-order induction can be an equilibrium of the meta-game. Similarly, first-order induction may be the selected equilibrium (as in Steiner and Stewart, 2008, Argenziano and Gilboa, 2012). Viewed thus, we suggest that second-order induction is a natural candidate for a focal point in the reasoning (meta-)game. Assuming that people do engage in this process in non-strategic set-ups, where it might lead to good predictions (as suggested by Propositions 2 and 4), we propose that in a strategic set-up second-order induction may be the equilibrium players coordinate on. Clearly, this is an empirical claim that needs to be tested. However, stopping at second-order induction doesn't not involve any assumption bounded rationality; it is only a specific theory of focal points in the reasoning game.

Lastly, we point out that higher orders of induction may generate identification problems: since the agents in our model are assumed to learn parameters (as the parameters of the similarity function in second-order induction), one should be concerned about higher orders of induction increasing the number of parameters. Surely, it is possible that third- or even fourth-order induction would be identifiable and generate better predictions. But an infinite regress is likely to generate a model that cannot be estimated from the finite database, and the optimal choice of the order of induction in the model may follow considerations such as the Akaike Information Criterion (Akaike, 1974).

4.6 Cases and Rules

As mentioned above, one can assume that people use rule-based, rather than case-based reasoning, and couch the discussion in the language of rules. Rules are naturally learnt from the data by a process of abduction (or case-to-rule induction), which can also be viewed as a type of second-order induction.

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 contra-

dict 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.

5 Appendix A: Proofs

Proof of Proposition 1:

Assume first that $j \in M$ is informative relative to $J \subseteq M \setminus \{j\}$ 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

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

It is straightforward that $\overline{y}_i^{(L,z')} - \overline{y}^{(L,z')} = O\left(\frac{1}{t}\right)$ and

$$MSE(L) - MSE'(L) = O\left(\frac{1}{t}\right). \tag{9}$$

Let us now consider the given set of variables J and $j \in M \setminus J$ that is informative relative to J. For any $z' \in \{0,1\}^J$ we have

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

and for z (for which $\overline{y}^{(J \cdot j, z \cdot 0)} \neq \overline{y}^{(J \cdot j, z \cdot 1)}$ is known),

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

where $c > 0$ is a constant that does not depend on t. It follows that

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

where $c' = \frac{|b(J,z)|}{n} c > 0$ is independent of t. This, combined with (9), means that $MSE(J \cup \{j\}) < MSE(J)$ 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 $MSE(J \cup \{j\}) = MSE(J)$. 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 $MSE'(J \cup \{j\}) = MSE'(J)$ 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 $MSE(L)$ are higher than those in $MSE'(L)$: 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 $MSE(J \cup \{j\}) > MSE(J)$. □

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 2:

Assume a data generating process (g, f) for which all $j \in M$ are informative. For a given $j \in M$ there exists $z^{-j} \in \{0,1\}^{m-1}$ such that $f\left(z^{-j} \cdot 0\right) \neq f\left(z^{-j} \cdot 1\right)$ (hence $\left[f\left(z^{-j} \cdot 0\right) - f\left(z^{-j} \cdot 1\right)\right]^2 = 1$) and $g\left(z^{-j} \cdot 0\right), g\left(z^{-j} \cdot 1\right) > 0$. Consider a proper subset of predictors, $J \subsetneq M$, and let $j \notin J$. Assume that n is large. Focus on an observation i whose x_i is in the bin defined by $z^{-j} \cdot 0$, and consider its estimated $\bar{y}_i$. In the computation of the latter (according to J, which does not include j) there are observations x_k in the bin defined by $z^{-j} \cdot 1$, and they contribute 1 to the sum of squared errors. Clearly, the opposite is true as well. Hence, focusing on these bins alone we find a lower bound of the sum of squared errors $\sum_{i=1}^{n} \left(\overline{y}_i^s - y_i\right)^2$ that is of the order of magnitude of $2ng\left(z^{-j} \cdot 0\right) g\left(z^{-j} \cdot 1\right)$. (We skip the standard approximation argument as in the proof of Proposition 1.)

For large enough n, we can therefore conclude that with arbitrarily high probability we have

$$MSE\left(J\right) - MSE\left(J \cup \{j\}\right) > g\left(z^{-j} \cdot 0\right) g\left(z^{-j} \cdot 1\right)$$

for every $J \subseteq M \backslash \{j\}$. Observe that there are finitely many bins, and therefore, for

a given $\delta > 0$ one can find N such that for every $n \geq N$

$$P\left(\begin{array}{c} MSE\left(J\right) - MSE\left(J \cup \{j\}\right) > g\left(z^{-j} \cdot 0\right) g\left(z^{-j} \cdot 1\right) \\ \forall j \in M, \forall J \subseteq M \backslash \{j\} \end{array} \right) \geq 1 - \delta. \qquad (10)$$

We now turn to select a value $\bar{c} > 0$ that would be small enough so that the reduction in the $AMSE$ thanks to omitting a variable j would not be worth the increase due to the error. For each j, let

$$d_j = \max \left\{ g\left(z^{-j} \cdot 0\right) g\left(z^{-j} \cdot 1\right) \;\middle|\; z^{-j} \in \{0,1\}^{m-1} \quad f\left(z^{-j} \cdot 0\right) \neq f\left(z^{-j} \cdot 1\right) \right\}$$

and

$$d \equiv \min_{j \in M} d_j.$$

Note that $d_j > 0$ for all j (as each j is informative), and hence $d > 0$. Set $\bar{c} = d/2$.

Given $\delta > 0$ let N be such that for every $n \geq N$ (10) holds. Let $c \in (0, \bar{c})$. We know that $MSE\left(M\right) = 0$ and $AMSE\left(M\right) = mc$. By the choice of $\bar{c}$, $\arg\min_{J \subseteq M} AMSE\left(J, c\right) = \{M\}$. Hence for any $\delta > 0$ there exists N such that for every $n \geq N$

$$P\left(\arg\min_{J \subseteq M} AMSE\left(J, c\right) = \{M\} \right) \geq 1 - \delta.$$

□

Proof of Proposition 3:

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 $MSE\left(J\right) = 0$ (and $AMSE\left(J\right) = c$). We will show that the proposition holds for J and J' that are (distinct) singletons.

Let the variables be generated according to the process described with $0 < \eta < 0.5$. Each x^j has a probability of equalling y that is at least η^n. The probability it does *not* provide a perfect fit is bounded above by $(1 - \eta^n) < 1$ – 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 $(1 - \eta^n)^m \to 0$ 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 $[1 - (1 - \eta^n)^m]^2 \to 1$ as $m \to \infty$. □

Proof of Theorem 1:

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 $MSE(J) \geq \frac{1}{n^3} > cm \geq R$, that is, $MSE(J) > R$ 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, $\{T_v\}_{v \in J}$ is a cover of P. It only remains to note that $AMSE(J,c) \leq R$ implies that $|J| \leq k$. □

Proof of Observation 2:

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 verse. (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. □

Proof of Proposition 4:

We wish to show that arbitrarily low values of the MSE can be obtained with probability that is arbitrarily close to 1. Let there be given $\nu > 0$ and $\xi > 0$. We wish to find N_0 and W_0 such that for every $n \geq N_0$, the vector w_0 defined by $w_0^j = W_0$ satisfies

$$P(MSE(w_0) < \nu) \geq 1 - \xi.$$

To this end, we first wish to 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 $N_0 \geq N_1$ such that, with probability of at least $(1 - \xi/2)$, for each $n \geq N_0$ 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.

Finally, we turn to choose W_0. For each i, the proportion of observations x_k with $[f(x_i) - f(x_k)]^2 > \nu$ is bounded above by $(1-\zeta)$. Define w_0 by $w_0^j = W_0$. Observe that, as $W_0 \to \infty$,

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

and this convergence is uniform in n (as the definition of ζ is independent of n). Thus a sufficiently high W_0 can be found so that, for all $n \geq N_0$, $MSE(w_0) < \nu$ with probability $(1-\xi)$ or higher. $\square$

Proof of Proposition 5:

The general idea of the proof is very similar to that of Proposition 3: non-uniqueness is obtained by showing that two variables can each provide perfect fit on their own. In the continuous case, however, to obtain perfect fit one needs a bit more than in the binary case: in the latter, it was sufficient to assume that there are at least two observations with $y_i = 0$ and two with $y_i = 1$; in the continuous case we need to make sure that each y_i has a close enough y_k. For this reason, we state and prove the result for a large n; yet, $M(n)$ will be larger still, so that we should think of this case as $m >> n$.

We now turn to prove the result formally. It will be convenient to define, for $w \in [0, \infty]^m$, $supp(w) = \left\{ l \in M \;\middle|\; w^l > 0 \right\}$.

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 $m \geq M(n)$,

$$P(\varepsilon\text{-}\arg\min AMSE \text{ is not connected}) \geq 1 - \delta.$$

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 $AMSE(w) \leq \varepsilon + c$.[27] For each j, 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 ε-arg min $AMSE$ is not connected, 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 2:

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 1. 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 1, and we use the same notation here. That is, we assume a given instance of SET-COVER: P, $r \geq 1$

[27] 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.

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 exists a vector w with $AMSE(w, c) \leq R$ iff a cover of size k exists for the given instance of SET-COVER.[28] 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 1, 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 1 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

[28] 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.

is no longer the case here. However, 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 $x^j_{2p+1} = x^j_{2p+2} = x^j_i (= 0)$ 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 are 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. □

References

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

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

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

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

[5] Argenziano, R. and I. Gilboa (2012), "History as a Coordination Device", *Theory and Decision*, **73**: 501-512.

[6] Argenziano, R. and I. Gilboa (2018), "Learning What is Similar: Precedents and Equilibrium Selection", working paper.

[7] Aumann, R. J. (1976), "Agreeing to Disagree", *The Annals of Statistics*, **4**: 1236-1239.

[8] Aumann, R. J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica*, **55**: 1-18.

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

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

[11] Brandenburger, A. and E. Dekel (1987), "Rationalizability and Correlated Equilibria", *Econometrica*, **55**: 1391-1402.

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

[13] Cortes, C., and V. Vapnik (1995), "Support-Vector Networks", *Machine Learning*, **20**: 273-297.

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

[15] ——— (1937), "La Prevision: ses Lois Logiques, ses Sources Subjectives", Annales de l'Institut Henri Poincare, **7**: 1-68.

[16] Eilat, R. (2007), "Computational Tractability of Searching for Optimal Regularities", working paper.

[17] 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.

[18] ——— (1952), "Discriminatory Analysis: Small Sample Performance". Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

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

[20] ——— (2001), *A Theory of Case-Based Decisions*, Cambridge: Cambridge University Press.

[21] ——— (2012), *Case-Based Predictions*. World Scientific Publishers, Economic Theory Series (Eric Maskin, Ed.), 2012.

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

[23] Gul, F. (1998), "A Comment on Aumann's Bayesian View", *Econometrica*, **66**: 923-928.

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

[25] Jaekel, F., B. Schoelkopf, and F. A. Wichmann (2008), "Generalization and Similarity in Exemplar Models of Categorization: Insights from Machine Learning", *Psychonomic Bulletin & Review*, **15**: 256-271.

[26] ——— (2009), "Does Cognitive Science Need Kernels?", *Trends in Cognitive Sciences*, **13**: 381-388.

[27] Kohavi, R. (1995), "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *Proceedings of the International Joint Conference on Artificial Intelligence*.

[28] Lieberman, O. (2010), "Asymptotic Theory for Empirical Similarity Models", *Econometric Theory*, **26**: 1032-1059.

[29] ——— (2012), "A Similarity-Based Approach to Time-Varying Coefficient Nonstationary Autoregression", *Journal of Time Series Analysis*, **33**: 484-502.

[30] Lieberman, O. and P. F. Phillips (2014), "Norming Rates and Limit Theory for Some Time-Varying Coefficient Autoregressions", *Journal of Time Series Analysis*, **35**: 592-623.

[31] ——— (2017), "A Multivariate Stochastic Unit Root Model with an Application to Derivative Pricing", *Journal of Econometrics*, **196**: 99-110.

[32] Medin. D. L. and M. M. Schaffer (1978), "Context Theory of Classification Learning", *Psychological Review*, **85**: 207-238.

[33] Morris, S. (1995), "The Common Prior Assumption in Economic Theory", *Economics and Philosophy*, **11**: 227-253.

[34] Nadaraya, E. A. (1964), "On Estimating Regression", *Theory of Probability and its Applications*, **9**: 141–142.

[35] Nosofsky, R. M. (1984), "Choice, Similarity, and the Context Theory of Classification", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**: 104-114.

[36] ——— (1988), "Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**: 700-708.

[37] ——— (2014), "The Generalized Context Model: An Exemplar Model of Classification", in *Formal Approaches in Categorization*, Cambridge University Press, New York, Chapter 2, 18-39.

[38] Park, B. U. and Marron, J. S. (1990), "Comparison of data-driven bandwidth selectors", *Journal of the American Statistical Association*, **85**: 66–72.

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

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

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

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

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

[44] Shepard, R. N. (1957), "Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space", *Psychometrika*, **22**: 325-345

[45] ——— (1987), "Towards a Universal Law of Generalization for Psychological Science", *Science*, **237**: 1317-1323

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

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

[48] Vapnik, V. (1998), *Statistical Learning Theory*, New York: John Wiley and Sons.

[49] ——— (2000), *The Nature of Statistical Learning Theory*, Berlin: Springer.

[50] Watson, G. S. (1964), "Smooth regression analysis", *Sankhyā: The Indian Journal of Statistics, Series A*, **26**: 359–372.

[51] Wittgenstein, L. (1922), *Tractatus Logico Philosophicus.* London: Routledge and Kegan Paul.