Precedents, Reputation, and Higher-Order Induction[*]

Rossella Argenziano[†] and Itzhak Gilboa[‡]

July 2017

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. Finally, we discuss more sophisticated forms of learning, by which similarity is defined not only on cases but also on attributes, and the importance of some attributes, learnt from the data by second-order induction, can also affect the perceived importance of other attributes.

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

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

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

1 Introduction

1.1 Motivating Example

The election of Obama as President of the US in 2008 was a defining event in US history. For the first time, a person who defines himself and is perceived by others as an African-American was elected for the highly coveted office, and this was clearly an important 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, and the statistics started looking differently.

It appears, however, that the single case of President Obama changes statistics far beyond its relative frequency, and this remains true even if we weigh cases by their recency. For example, considering only the post-WWII period, the US had 11 presidents before Obama. The effect of his election, however, does not seem to be captured by the difference between 0:11 and 1:12. We suggest that 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: first, to estimate the probabilities of future events according to the relative frequency of similar events in the past, and, second, to learn what counts as “similar”. Up to Obama’s election, “race” was an important attribute, one which could be a useful variables in fitting the data. An African-American was therefore considered dissimilar to all past presidents, differing from them on an attribute that proved to have predictive power. 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 the data. And if second-order induction suggests that similarity between cases can be more accurately judged ignoring this variable, it is easier to understand the importance of the precedent.

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. Statistical techniques are used to predict election results based on polls in the days or weeks preceding an election, but such techniques can hardly be used to provide reliable probability estimates for the success of a candidate who considers running, say, two years in advance.

Yet, people are faced with decision problems that depend on these probabilities. The potential candidate herself has to decide whether to run for office, a decision involving very high stakes. Potential donors and volunteers also ask themselves how likely it is that this candidate should win. These beliefs are usually not estimated in any scientific or objective way, and often not even in any explicit way. Moreover, one may argue that events of this type do not have well-defined probabilities. But, in the final analysis, decisions have to be made, and some form of beliefs, probabilistic or not, explicit or not, would underlie rational agents' choices. 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. 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. 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 both in statistics as "kernel methods" (Akaike, 1954, Rosenblatt, 1956, Parzen, 1962) and in psychology as "exemplar learning" (see Medin and Schaffer, 1978, and Nosofsky, 1988). Gilboa, Lieberman, and Schmeidler (GLS, 2006) suggested the notion of learning the similarity function from the data, and referring to the optimal function as the "empirical similarity". While their paper can be viewed as suggesting a statistical technique, our focus in this paper is on the interpretation of the model as a description of the way people reason. Consider, for example, the choice of an electronic device such as a smartphone or a computer. A child, or an inexperienced adult, may assume that products with similar outward appearance would be of similar quality. Experienced consumers, however, would know that the name of the producer counts more than, say, screen size, when it comes to quality. We view this as an example of second-order induction. Clearly, consumers do not explicitly compute the empirical similarity in as GLS (2006); but the model is proposed as an idealized account of a process that people do engage in, typically implicitly.

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 computation, 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" in the example 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 then show how these results can be applied to the questions of (i) the importance of precedents; and (ii) the cost of establishing and retaining reputation.

The rest of the paper is organized as follows. Section 2 presents the basic model and the idea of the empirical similarity formally. Section 3 offers a few general results, whereas Section 4 interprets them for the analysis of precedents and of reputation. Finally, Section 5 concludes with a general discussion.

2 Case-Based Beliefs

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 $s : X \times X \rightarrow$

$\mathbb{R}_+$, such that the probability that $y_p = 1$ is estimated by

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

if $\sum_{i \leq n} s(x_i, x_p) > 0$ and $\overline{y}_p^s = 0.5$ otherwise. [1]

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

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

Thus, 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.

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: the "empirical similarity" is a similarity function that, had it been used to predict the existing data points, where each is estimated based on the others, it would have performed well. That is, 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

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

and consider the sum of squared errors,

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

[1] 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.

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

The focus of our analysis is the question, whether a variable is important for prediction or not. 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 $SSE\left(J \cup \{j\}\right)$. If $SSE\left(J\right) > SSE\left(J \cup \{j\}\right)$, then the inclusion of the variable j provides a better fit to the data, and the variable will be used for future predictions. If, by contrast, $SSE\left(J\right) < SSE\left(J \cup \{j\}\right)$, the addition of the variable j results in higher errors, and it will not be included. Intuitively, we can think of the variable as "adding noise". The reason that a variable can decrease the SSE is related to "the curse of dimensionality": a set of predictors J splits the database into "bins", namely, sub-databases with identical $\left(x^l\right)_{l \in J}$ values. A new variable splits each of these bins into smaller ones, so that the number of bins grows exponentially in $|J|$. When there are too few observations in a bin, the prediction error can grow. Intuitively, if we are too picky about the notion of similarity, there will not be enough similar cases for any given case. Note that this reason is distinct from overfitting, which may be yet another reason to prefer small sets of predictors.

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, $SSE\left(J\right) = SSE\left(J \cup \{j\}\right)$ 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.

3 Results

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.[2] 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.

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
$y = 1$	K	k

Thus, the database includes $L + K$ cases with $x^j = 0$, of which in L we have $y = 0$, and in the other K – the value $y = 1$ was observed; and it also includes additional $k + l$ cases with $x^j = 1$, of which l have $y = 0$ and k have $y = 1$.

We are interested in the sign of

$$\Delta\left(K, L, k, l\right) \equiv SSE\left(\{j\}\right) - SSE\left(\varnothing\right)$$

Clearly, $\Delta\left(K, L, k, l\right) = \Delta\left(L, K, l, k\right)$ and $\Delta\left(K, L, k, l\right) = \Delta\left(k, l, K, L\right)$, as the SSE calculations do not change if we switch between 0 and 1 either for a predictor x^j or for the predicted variables y.

Notice that $\Delta\left(K, L, k, l\right) > 0$ implies that the variable j is not included in the empirical similarity function, whereas $\Delta\left(K, L, k, l\right) < 0$ implies that it is. Of particular interest would be cases that *change* the sign of Δ, for example,

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

$\Delta(K, L, k, l) > 0$ but $\Delta(K, L, k+1, l) < 0$, where a single case with $x^j = 1$ and $y = 1$ adds the variable j to the empirical similarity function, or, vice versa, if $\Delta(K, L, k, l) < 0$ but $\Delta(K, L, k+1, l) > 0$ so that a single case as above induces the omission of j, rendering a previously-important variable unimportant.

When would the variable x^j be informative enough to render $\Delta(K, L, k, l)$ negative? 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 $K/L = k/l$, there is no predictive power to be gained from splitting the database according to x^j, and one would expect to find $\Delta(K, L, k, l) > 0$ (where the inequality follows from the loss of accuracy when using smaller sub-databases). If, by contrast, $K/L \neq k/l$, then x^j provides statistical information about y. Whether the additional statistical information is worth splitting the database into smaller bins would depend on the sizes of the bins obtained. Proposition 1 in Argenziano and Gilboa (2017) considers the case in which a database is replicated, that is, the matrix

$$\begin{array}{lll}
\#\text{ of cases} & x^j = 0 & x^j = 1 \\
y = 0 & tL & tl \\
y = 1 & tK & tk
\end{array}$$

for $t > 0$. It implies that (in this rather special case) if, indeed, $K/L = k/l$, then $\Delta(tK, tL, tk, tl) > 0$ for all t. By contrast, if $K/L \neq k/l$, then $\Delta(tK, tL, tk, tl) < 0$ for sufficiently large t.

Our focus in this paper is, however, on databases for which k and l are 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 k and l as small (and sometimes zero). Moreover, in many of these cases the number of relevant cases in the entire history isn't very large either. For example, the number of presidential campaigns that can be considered sufficiently relevant to a given US election will be in the dozens, rather than the thousands. Hence, here we have limited interest in asymptotic results.

The case $k = l = 0$ will be of special interest. It 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.

We assume that there is a non-trivial history in which $x^j = 0$. Specifically, we assume throughout that $L, K > 2$. 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 $y = 1$.

The first result we wish to establish is that, if there are regularities in the database, the empirical similarity will spot them. Intuitively, we would like to say that, if it so happens that *all* cases in the database with $x^j = 1$ had the same y value, then the variable j will be included in the empirical similarity function, and will be perceived to be of predictive power. This statement need not hold if there is only one case with $x^j = 1$ in the database. But if "*all* cases in the database with $x^j = 1$" refers to at least two such cases, the result holds true. Formally,

Proposition 1 For any (K, L), and any $k, l > 1$, we have

$$\Delta(K, L, k, 0) \ , \ \Delta(K, L, 0, l) < 0.$$

Recall that we assume that $L, K > 2$, so that the sub-database for which $x^j = 0$ does not suggest a clear regularity about y. By contrast, if we focus, for instance, on the case $k > 1$,$l = 0$, the rule "if $x^j = 1$ then $y = 1$" holds in the database – where its antecedent is satisfied at least twice. Under these conditions, the empirical similarity will "identify" the rule by including the variable j in the 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 implicitly learn regularities. If it is indeed the case that $x^j = 1$ implies $y = 1$, the empirical similarity will "learn" the regularity by highlighting the importance of x^j. Clearly, the variable might be found important also when no such simple regularity can be found. Indeed, case-based predictions can prove useful when no simple rules hold true. But it is reassuring to know that, should such rules exist, they will not be missed by a case-based reasoner who employs second-order induction.

The parameter values $k = 1, l = 0$ (or vice versa, $k = 0, l = 1$) are not covered by Proposition 1. They might appear to represent a relatively degenerate and uninteresting special case. However, in the next section we will discuss applications where these values appear to be in the limelight. They correspond to new realities, where $x^j = 1$ has never been observed before and therefore deserve analysis. It turns out 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 the corresponding y value was the less frequent value in the rest of the database. Formally,

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

We find this result rather intuitive: when no cases with $x^j = 1$ were ever observed ($k = l = 0$), there is no real meaning to the variable x^j: it is always 0 and can be ignored.[3] 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

[3] 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.

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.

A complete classification of the quadruples (K, L, k, l) for which $\Delta(K, L, k, l)$ is negative, positive, or zero is beyond the scope of this paper. However, a few intuitively interpretable results can be obtained when $K = L$, that is, if the database with $x^j = 0$ is evenly split between $y = 0$ and $y = 1$:

Proposition 3 Let $K = L > 2$, $l \geq k > 0$.[4] Define $w = l - k \geq 0$. The following hold:

(i) For every k, $\Delta(L, L, k, k)$,$\Delta(L, L, k, k+1) > 0$.

(ii) For every k, $\Delta(L, L, k, k+w)$ is decreasing in w.

(iii) For every k and every $w \geq 2$, $\Delta(L, L, k, k+w)$ is increasing in k.

(iv) For every k, there exists $w(k) \geq 2$ such that

$$\begin{aligned} \forall w &< w(k) \qquad \Delta(L, L, k, k+w) \geq 0 \\ \forall w &\geq w(k) \qquad \Delta(L, L, k, k+w) < 0. \end{aligned}$$

(v) $w(k)$ is non-decreasing in k.

This proposition analyzes the behavior of the function $\Delta(L, L, k, l)$ as k and/or l increase.[5] Notice that, when $l > 1$, $\Delta(L, L, 0, l)$ is known to be negative by Proposition 1. This means that the variable x^j is used for prediction, because throughout the database, when $x^j = 1$ it was also true

[4] The case $k \geq l > 0$ clearly has symmetrical properties.

[5] Note that, due to symmetry, $\Delta(L, L, k, l) = \Delta(L, L, l, k)$.

that $y = 0$. Let us now consider the same function as k grows. Part (i) says that, as k becomes equal to $(l-1)$ or to l, $\Delta(L, L, k, l)$ is positive, so that x^j is no longer used for prediction. Intuitively, when there are sufficiently many (k) counter-examples to the rule "$x^j = 1$ implies $y = 0$", the variable is considered unimportant. Part (iv), however, says that if k keeps growing (holding L and l fixed), the variable will re-appears in the empirical similarity function. Notice that in this time it would be considered informative not because $x^j = 1$ is associated with $y = 0$ (as for the case $k = 0$), but for the opposite association. Indeed, when there are going to be many cases in which $(x^j = 1, y = 1)$, eventually the l cases in which $(x^j = 1, y = 0)$ were observed will be implicitly considered random errors. To sum, Proposition 3 shows that, when $K = L$, for every $l > 1$, the variable x^j will be used for prediction for very low and very high values of k, but not for some intermediate values.

4 Applications

4.1 Precedents

We suggest to interpret Proposition 3 as capturing the way that a precedent makes a variable lose importance. Consider our motivating example, namely, the election of President Obama. We focus on the variable x^j which denotes race, where $x^j = 1$ means that the candidate is African-American. The database would include cases of people who ran campaigns to become presidents, at least in the primaries of their party, or made a similar attempt to be elected. The vast majority of them were white, namely, had $x^j = 0$. Assume that, on top of these white candidates (of which L failed, and K succeeded, where the proposition requires $K = L$), there also some attempts made by African-American candidates, but all of those failed. Assume that the number of these attempts, l, is at least two. Given zero successes by African-American candidates, $k = 0$, race would seem to be important (by Proposition 1), and the similarity function would take it into account, implicitly noticing the regularity “No African-American was ever elected president”. However, when a few exceptions exist ($k = l - 1$ would suffice), Part (i) of Proposition 3 will imply that race becomes unimportant. This is the sense in which the precedent changes more than the statistics: it changes the similarity function as well. If we let k grow, that is, if after Obama all future cases are of successful African-American candidates, at some point race will be considered important again, this time because $x^j = 1$ is predicts victory.

Note that Proposition 3 is limited to the case $K = L$. This assumption doesn’t seem realistic: if we wish to argue that all candidates who ran for the party’s nomination are considered to be cases, then, more or less by definition, $L > K$. Moreover, the account above, and Proposition 3 do not guarantee that a *single* precedent would be enough to render the race variable unimportant. Unfortunately, we cannot offer any general results about the case $L > K$. However, direct computation shows that, for many values of the parameters, with $L > K$ and $l = 2$, the value $k = 1$ is enough to obtain $\Delta(K, L, k, l) > 0$.

4.2 Reputation

Consider an agent who’s new to an economic or political scene, and who’s trying to establish reputation. For example, we may consider a new dean who aims to enforce regulations more strictly than her predecessors, or a central banker who intends to curb inflation. Assume that the variable x^j is the agent’s proper name, so that, starting with a clean slate, there are no cases with $x^j = 1$. Let us assume, again for simplicity, that in the database past agents who assumed the same post had an equal number of successes and failures, so that $K = L$ and $k = l = 0$.

Propositions 2 and 1 suggest that the new agent would have to invest an effort in establishing $y = 1$ twice in order to establish her reputation: at the outset, with $k = l = 0$, there are no cases with $x^j = 1$, and the variable clearly does not aid in prediction. But even with $k = 1$, $\Delta(K, L, 1, 0) > 0$, and x^j does not enter the empirical similarity function. However, with $k = 2$ it does. That is, the dean who wishes to convey the message that "the rules have changed" would have to successfully enforce the rules twice.

Proposition 3 also tells us what will happen if the dean fails to enforce the rules at the beginning of her tenure. Part (iv) of the proposition guarantees that for any number l of failures, it would still be possible to establish reputation for implementing $y = 1$ by having a large enough number of successes k. Part (v) points out that, while establishing reputation after allowing some failures is possible, the cost of doing so will increase. Thus, second-order induction can explain why it is easier to establish reputation given a clean slate than it is to re-establish it after some failures.

5 Discussion

5.1 Additional Examples

5.1.1 Example: The Collapse of the USSR

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

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

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

5.1.2 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.

5.2 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.[6]

5.3 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 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

[6] 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.

the weight of another variable.

To consider another example, let us revisit the example of the collapse of the USSR (5.1.1 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.

5.4 Compatibility with Common Knowledge of Equilibrium Selection

The informal discussions of precedents and of reputation implicitly assumes that a large population of players is involved in a coordination game, and that the equilibrium selection is done by the prediction of play using the empirical similarity. We argue in Argenziano and Gilboa (2017) that computing the empirical similarity is compatible with a Bayesian approach applied to the grand state space; but is it also compatible with common knowledge of the rationality of others, and of the model itself?

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

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

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

Along similar lines, in our case there may be some players who are non-strategic and do indeed believe that the process follows an unknown similarity function. Attempting to estimate this function, they would use the empirical similarity to generate their prediction over the game's equilibrium. Other players might engage in Level-1 reasoning, and optimally react to the existence of Level-0 players by predicting the equilibrium chosen by the empirical similarity. As this is a coordination game, best response implies behavior according to the Level-0 beliefs. Similarly, higher levels of reasoning would also follow the same equilibrium prediction. In other words, the Level-0 prediction, which is statistically but not strategically sophisticated, isn't only a reasonable focal point in the belief-selection equilibrium; it is also the best prediction of the strategic choices of players who engage in Level-k reasoning for any $k \geq 0$ (see Stahl and Wilson, 1995, Nagel, 1995).

6 Appendix: Proofs

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

Proof of Proposition 1:

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

The SSE's are given by

$$\begin{aligned} SSE(\varnothing) &= K\left(1 - \frac{K-1}{L+K+l-1}\right)^2 + (L+l)\left(\frac{K}{L+K+l-1}\right)^2 \\ &= K(L+l)\frac{L+l+K}{(L+l+K-1)^2} \end{aligned}$$

and

$$\begin{aligned} SSE(\{j\}) &= K\left(1 - \frac{K-1}{L+K-1}\right)^2 + L\left(\frac{K}{L+K-1}\right)^2 \\ &= LK\frac{L+K}{(L+K-1)^2} \end{aligned}$$

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

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

which is clearly negative. Notice that the above holds for any $L, K > 2$, with no assumptions made about their relative sizes. □

Proof of Proposition 2:

We need to show that

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

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

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

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

The SSE's are given by

$$\begin{aligned} SSE(\varnothing) &= K\left(1-\frac{K-1}{L+K}\right)^2 + (L+1)\left(\frac{K}{L+K}\right)^2 \\ &= K(L+1)\frac{L+K+1}{(L+K)^2} \end{aligned}$$

and

$$\begin{aligned} SSE(\{j\}) &= K\left(1-\frac{K-1}{L+K-1}\right)^2 + L\left(\frac{K}{L+K-1}\right)^2 + 0.25 \\ &= LK\frac{L+K}{(L+K-1)^2} + 0.25 \end{aligned}$$

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

It follows that

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

Let $a(K,L)$ denote the the expression in the square brackets in (the RHS of) equation (3), which clearly has the same sign as $\Delta(K,L,0,1)$. First, we observe that

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

This establishes Part (iii), and will also be a useful benchmark for Parts (i) and (ii). Indeed, to prove that $a(K,L) > 0$ (and thus that $\Delta(K,L,0,1) > 0$) for $L > K$, we will consider the partial derivative of $a(K,L)$ relative to its second argument, and show that it is positive for $L \geq K$. (Clearly, $a(K,L)$

is a polynomial in its two arguments, and it is well-defined and smooth for all real values of (K, L).) To see this, observe that

$$\begin{aligned}\frac{\partial a\left(K,L\right)}{\partial L} &= 4L^3+12L^2K-6L^2+4LK^2+4LK+2L-4K^3+6K^2+2K \quad (4)\\ &= 4L^3+\left(12K-6\right)L^2+\left(4K^2+4K+2\right)L+\left(-4K^3+6K^2+2K\right)\end{aligned}$$

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

We now turn to the case $L<K$, where equation (4) 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 $K=L+1$. Here direct calculations yield

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

This time we consider the partial derivative of $a\left(K,L\right)$ wrt to its first 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 $a\left(K,L\right)$, and therefore the negativity of $a\left(K,L\right)$ (and of $\Delta\left(K,L,0,1\right)$) for $K>L$ would be established.

Consider, then,

$$\begin{aligned}\frac{\partial a\left(K,L\right)}{\partial K} &= 4L^3+4L^2K+2L^2-12LK^2+12LK+2L-12K^3+6K^2+10K-4 \quad (5)\\ &= 4L^3+\left(4K+2\right)L^2+\left(12K-12K^2+2\right)L+\left(6K^2-12K^3+10K-4\right)\end{aligned}$$

Consider (5) as a polynomial in L. Using the fact $L<K$, we obtain an upper bound on this expression by replacing L by K whenever coefficients of

powers of L are clearly positive:

$$\begin{aligned} & 4L^3 + (4K+2)L^2 + \left(12K - 12K^2 + 2\right)L + \left(6K^2 - 12K^3 + 10K - 4\right) \\ < \; & 4K^3 + (4K+2)K^2 + 12K^2 + 2K - 12K^2L + 6K^2 - 12K^3 + 10K - 4 \end{aligned}$$

and, if we drop the only term still involving L $(-12K^2L)$, we obtain

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

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

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

$$\begin{aligned} a(4,3) &= -204 \\ a(5,3) &= -1,424 \\ a(5,4) &= -496 \end{aligned}$$

This concludes the proof of Parts (i) and (ii). □

Proof of Proposition 3:

We start by noting that the SSE formulae, for the case $K = L$ and $l = k + w$, are

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

$$\begin{aligned} SSE(\{j\}) &= L\left(1 - \frac{L-1}{2L-1}\right)^2 + L\left(\frac{L}{2L-1}\right)^2 \\ &\quad + k\left(1 - \frac{k-1}{2k+w-1}\right)^2 + (k+w)\left(\frac{k}{2k+w-1}\right)^2 \end{aligned}$$

and thus

$$\begin{aligned}\Delta(L,L,k,k+w) &= 2\frac{L^3}{(2L-1)^2}+k(2k+w)\frac{k+w}{(2k+w-1)^2} \\ &\quad -(L+k)(2L+2k+w)\frac{L+k+w}{(2L+2k+w-1)^2}\end{aligned} \tag{6}$$

(i) To see that $\Delta(L,L,k,k)>0$, note that

$$\begin{aligned}&\Delta(L,L,k,k) \\ = {}& 2\left(\frac{L^3}{(2L-1)^2}+\frac{k^3}{(2k-1)^2}-\frac{(L+k)^3}{(2L+2k-1)^2}\right) \\ = {}& \frac{2}{(2L-1)^2(2k-1)^2(2L+2k-1)^2}\left[\begin{array}{c} L^3(2k-1)^2(2L+2k-1)^2 \\ +k^3(2L-1)^2(2L+2k-1)^2 \\ -(L^3+3L^2k+3Lk^2+k^3)(2L-1)^2(2k-1)^2\end{array}\right] \\ = {}& \frac{2Lk}{(2L-1)^2(2k-1)^2(2L+2k-1)^2}\left[\begin{array}{c} L^3(16k-4)+L^2(8(4k-1)(k-1)) \\ +L(12k^3-40k^2+24k-3) \\ +(4Lk^3-4k^3)+8k^2-3k\end{array}\right]>0\end{aligned}$$

where the inequality holds because the expression in square brackets can have only one term that may be negative, namely $L(12k^3-40k^2+24k-3)$, which is negative for $k=1,2$. However, it can easily be verified that for these two values the entire expression in square brackets is, indeed, positive.

We now turn to evaluate $\Delta(L,L,k,k+1)$. Here we simply note that

$$\Delta(L,L,k,k+1)=\frac{1}{4}L\frac{2k^2(4L-1)+2kL(4L-1)+(2L-1)^2}{k(2L-1)^2(L+k)}>0$$

which concludes the proof of Part (i). □

(ii) We wish to show that $\Delta(L,L,k,k+w)$ is decreasing in w. Consider Equation (6) and differentiate wrt w to obtain

$$\begin{aligned}&\frac{\partial\Delta(L,L,k,k+w)}{\partial w} \\ = {}& -\frac{Lz(L,k,w)}{(2k+w-1)^3(2L+2k+w-1)^3}\end{aligned} \tag{7}$$

where

$$\begin{aligned} z\left(L,k,w\right) &= w^4\left[L+2k-2\right] \\ &\quad +w^3\left[2L^2+6L\left(2k-1\right)+12k\left(k-1\right)+6\right] \\ &\quad +w^2\left[\begin{array}{c} L^2\left(12k-6\right)-6+ \\ 12L\left(-2k+3k^2+1\right)+24k\left(-k+k^2+1\right) \end{array}\right] \\ &\quad +w\left[\begin{array}{c} L^2\left(16k^2-8k+6\right)-20k \\ -L\left(-32k^3+24k^2-36k+10\right) \\ +36k^2-16k^3+16k^4+2 \end{array}\right] \\ &\quad +\left[\begin{array}{c} L^2\left(12k-2\right)+L\left(36k^2-24k+3\right) \\ +\left(24k^3-24k^2+6k\right) \end{array}\right] \end{aligned} \tag{8}$$

As the denominator in (2) is positive, we need to show that so it $z\left(L,k,w\right)$ for $w>0$. To prove this, we will argue that all the coefficients of powers of w in the above (Equation (8)) are positive. Specifically, the coefficients of w^4, w^3, w^2 and the free coefficient can be directly verified to be positive:

$$\begin{aligned} L+2k-2 &> 0 \\ 2L^2+6L\left(2k-1\right)+12k\left(k-1\right)+6 &> 0 \\ L^2\left(12k-6\right)-6+12L\left(-2k+3k^2+1\right)+24k\left(-k+k^2+1\right) &> 0 \\ L^2\left(12k-2\right)+L\left(36k^2-24k+3\right)+\left(24k^3-24k^2+6k\right) &> 0. \end{aligned}$$

As for the coefficient of w,

$$\begin{aligned} & L^2\left(16k^2-8k+6\right) \\ & -L\left(-32k^3+24k^2-36k+10\right) \\ & +\left(16k^4-16k^3+36k^2-20k+2\right) \end{aligned}$$

it is easy to see that the first and last terms are positive. Finally, straightforward analysis shows that $(-32k^3+24k^2-36k+10)$ is negative for all $k>0$. □

(iii) We need to show that, for every k and every $w\geq 2$, $\Delta\left(L,L,k,k+w\right)$ is increasing in k. Differentiating $\Delta\left(L,L,k,k+w\right)$ with respect to k we obtain

$$\frac{\partial\left(\Delta\left(L,L,k,k+w\right)\right)}{\partial k}=$$
$$\frac{2L}{\left(2k+w-1\right)^{3}\left(2L+2k+w-1\right)^{3}}P\left(L,k,w\right)$$

where

$$\begin{aligned}
P\left(L,k,w\right) &= w^{5}\\
&+w^{4}\left[3L+6k\right]\\
&+w^{3}\left[2L^{2}+12Lk+12k^{2}-6\right]\\
&+w^{2}\left[L^{2}\left(4k+2\right)+L\left(12k^{2}-12\right)+8k^{3}-24k+8\right]\\
&-w\left[L\left(36k-12\right)+k\left(36k-24\right)+6L^{2}+3\right]\\
&+\left[24k^{2}-L^{2}\left(12k-2\right)-L\left(36k^{2}-24k+3\right)-6k-24k^{3}\right]
\end{aligned}$$

We argue that $P\left(L,k,w\right)>0$ for all $w\geq 2$. We first observe that $P\left(L,k,2\right)>0$:

$$\begin{aligned}
P\left(L,k,2\right) &= 21L+42k+12Lk^{2}+4L^{2}k\\
&+48Lk+14L^{2}+48k^{2}+8k^{3}+10
\end{aligned}$$

which is clearly positive. Next, to see that $P\left(L,k,w\right)>0$ also for all $w>2$, we consider the derivative of $P\left(L,k,w\right)$ with respect to w and show that it is positive. Indeed,

$$\begin{aligned}
\frac{\partial P\left(L,k,w\right)}{\partial w} &= 5w^{4} \qquad (9)\\
&+w^{3}\left[12L+24k\right]\\
&+w^{2}\left[6L^{2}+36Lk+36k^{2}-18\right]\\
&+w\left[8L^{2}k+4L^{2}+24Lk^{2}-24L+16k^{3}-48k+16\right]\\
&+\left[-6L^{2}-36Lk+12L-36k^{2}+24k-3\right]
\end{aligned}$$

As the coefficients of w^{4},w^{3},w^{2}, and w in (9) are all positive, we can bound

$\frac{\partial P(L,k,w)}{\partial w}$ from below by $\frac{\partial P(L,k,w)}{\partial w}|_{w=1}$ and observe that

$$\begin{aligned}\frac{\partial P(L,k,1)}{\partial w} &= 5+[12L+24k]+\left[6L^2+36Lk+36k^2-18\right] \\ &\quad+\left[8L^2k+4L^2+24Lk^2-24L+16k^3-48k+16\right] \\ &\quad+\left[-6L^2-36Lk+12L-36k^2+24k-3\right] \\ &= 8L^2k+4L^2+24Lk^2+16k^3>0 \qquad (10)\end{aligned}$$

Because $P(L,k,w)>0$ was shown to hold at $w=2$, $\frac{\partial P(L,k,1)}{\partial w}>0$ for all $w\geq 1$, we conclude that $P(L,k,w)>0$ holds for all $w\geq 2$. □

(iv) Let there be given $k>1$. We wish to show that there exists $w(k)\geq 2$ such that

$$\begin{aligned}\forall w &< w(k) \qquad \Delta(L,L,k,k+w)\geq 0 \\ \forall w &\geq w(k) \qquad \Delta(L,L,k,k+w)<0.\end{aligned}$$

Given Part (i), we know that $\Delta(L,L,k,k+w)>0$ holds for $w=0,1$. Moreover, Part (ii) established that $\Delta(L,L,k,k+w)$ is strictly decreasing in w. Given these, it suffices to show that *for some* w, we have $\Delta(L,L,k,k+w)<0$, and we can then define $w(k)$ as the minimal w for which this inequality holds. We now turn to show that $\Delta(L,L,k,k+w)$ becomes negative as $w\to\infty$.

To see this, we consider the SSE expressions again:

$$\begin{aligned}SSE(\varnothing) &= \frac{(L+k)(L+k+w)^2+(L+k+w)(L+k)^2}{(2L+2k+w-1)^2} \\ SSE(\{j\}) &= \frac{2L^3}{(2L-1)^2}+k(k+w)\frac{2k+w}{(2k+w-1)^2}\end{aligned}$$

For any L and k, letting $w\to\infty$, we have

$$\begin{aligned}SSE(\varnothing) &= \frac{(L+k)(L+k+w)^2+(L+k+w)(L+k)^2}{(2L+2k+w-1)^2} \\ &= \frac{(L+k)w^2+...}{w^2+...}\to_{w\to\infty} L+k\end{aligned}$$

while

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

for $L > 2$ and

$$\frac{k(k+w)\,(2k+w)}{(2k+w-1)^2} \rightarrow_{w\rightarrow\infty} k$$

hence, $SSE\,(\varnothing) > SSE\,(\{j\})$ for all w large enough. □

(v) We wish to show that $w\,(k)$ is increasing in k. As noted above, Part (i) establishes that $w\,(k) \geq 2$ for all k. Consider $k, k' > 1$ with $k' > k$. We need to show that $w\,(k') \geq w\,(k)$. This, however, follows from Parts (iii) and (iv): by definition of $w\,(k')$, we have

$$\Delta\,(L, L, k', k' + w\,(k')) < 0.$$

Since $w\,(k') \geq 2$, we can apply Part (iii) to conclude that – given that $k < k'$ – we have

$$\Delta\,(L, L, k, k + w\,(k')) < \Delta\,(L, L, k', k' + w\,(k')) < 0$$

and this implies that $w\,(k) \leq w\,(k')$. □□

7 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), "Second-Order Induction and Agreement", *working paper.*

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

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. (1988), "Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality", *Journal of Experimental Psychology*, **4**: 700-708.

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.

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

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

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.