# 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 $\alpha$ 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

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

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

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

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 $4LW^{2}-4W^{3}\geq0$ 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 $\Delta$ 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 $\Delta$ 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 $\Delta$ 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 $\Delta$ 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 $\omega$ 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 $\omega$, $b(L, W, l, \omega)$ 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 $\omega$ 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 $\omega$). It follows that the smallest integer $\omega$ ($\omega > 1$) 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 $\omega$, 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 $\omega$ for $\omega\geq 1$.

**Proof:**

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

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

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 $\omega$ 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 $\omega$ on the numerator of (9) is strictly positive for positive values of $\omega$. Hence, if we can show that the polynomial is positive for some positive value of $\omega$, then it is positive for all larger values of $\omega$ 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 $\omega$ large enough. Hence it has to have a root at some $\omega > 1.5$. Further, it is unique because $b$ is strictly decreasing in $\omega$ 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 $\omega$ with a fourth degree polynomial (in $\omega$) 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 $\Delta$ 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 $w>-\frac{W}{L}$.