PRIOR SYMMETRY, SIMILARITY-BASED REASONING, AND ENDOGENOUS CATEGORIZATION*

MARCIN PĘSKI

ABSTRACT. This paper presents a rational theory of categorization and similarity-based reasoning. I study a model of sequential learning in which the decision maker infers unknown properties of an object from information about other objects. The decision maker may use the following heuristics: divide objects into categories with similar properties and predict that a member of a category has a property if some other member of this category has this property. The environment is *symmetric*: the decision maker has no reason to believe that the objects and properties are a priori different. In symmetric environments, categorization is an optimal solution to an inductive inference problem. Any optimal solution looks *as if* the decision maker categorizes. Various experimental observations about similarity-based reasoning coincide with the optimal behavior in my model.

1. INTRODUCTION

In a class of discrete prediction problems, the decision maker (henceforth called the DM) is supposed to predict whether an object $o \in O$ has a property $p \in P$. Quite often, the prediction is based on some notion of similarity: the DM predicts that o has property p if a similar object o' has this property. Table 1 presents an example of such reasoning. The DM uses information about chickens and tigers to predict that a falcon, which is more similar to chicken than to a tiger, does not mix oxygen with carbon dioxide while breathing. Arguably, the knowledge that a chicken is similar to a falcon is not inborn to the DM, but arises as a result of learning that both correspond to objects with wings, feathers and beaks and share other properties. From this perspective, similarity-based reasoning is an application of the Principle of Similarity: a presumption that similarity of some properties of two objects

This paper was previously presented under the title "Categorization." I am grateful to Jeff Ely, Sham Kakade, Philip Reny, John Conlon, Willie Fuchs, Ali Hortacsu, Motty Perry, Bart Lipman, Eddie Dekel, Alberto Bisin and Jorg Stoye for helpful conversations and participants at seminars at the University of Toronto, University of Chicago, MIT/Harvard Theory Seminar, Center of Rationality, Tel Aviv University and New York University for discussion and comments. P. Reny and the referee suggested simplifications in the statement and the proof of Theorem 1. I am grateful to the editor and the referee for numerous and helpful suggestions. The author bears full responsibility for the remaining errors.

Department of Economics, University of Texas at Austin, 1 University Station #C3100, Austin, Texas 78712. *E-mail*: mpeski@gmail.com.

Premise:	Chicken does not mix oxygen with carbon dioxide while breathing
	Tiger mixes oxygen with carbon dioxide while breathing
Conclusion:	Falcon does not mix oxygen with carbon dioxide while breathing

TABLE 1. Example of similarity-based reasoning.

indicated the similarity of other properties. This principle, in various forms, is recognized as one of the most salient features of human reasoning and the foundation of any inductive argument (for example, Mill (1874), Hume (1777)). The principle underlies automatized induction in so-called collaborative filtering problems.

Categorization is one of the models proposed to capture properties of similarity-based reasoning (for various other models, see Tversky (1977), Gilboa and Schmeidler (1995) or Osherson, Wilkie, Smith, Lopez, and Shafir (1990)). There are two defining characteristics of categorization. First, the DM divides objects and properties into a finite number of groups called *categories* and considers two objects to be similar if they are assigned to the same category. Thus, similarity has a particularly simple binary form. Second, prediction is category-based: all members of each category are presumed to share properties, and an observation that a member of a category has a certain property is generalized to the other members. In the above example, the DM's reasoning can be interpreted as predicting a falcon's properties given that it falls into the category "birds."

In the paper, I argue that both similarity-based reasoning and endogenous categorization have a rational explanation as optimal behavior in a model of sequential learning. There are infinite sets of objects O and properties P. Each period t, the DM is asked whether object $o_t \in O$ has property $p_t \in P$. After making a prediction, the DM learns the correct answer $\theta(o_t, p_t) \in \{0,1\}$, where function $\theta : O \times P \to \{0,1\}$ is called the *state of the world*. While making a prediction, the DM uses the information that she acquired in the previous periods.

The key assumption is the symmetry of the distribution ω from which state of the world θ is drawn. The assumption says that the prior is invariant with respect to relabelling of objects and properties. This means that a priori all objects and properties are considered by the DM to be perfectly symmetric and exchangeable. In particular, the assumption eliminates any exogenous categorization before the DM observes any data.

The symmetry assumption has a sound interpretation from the subjective point of view, where ω is treated as the beliefs of the Bayesian DM. The symmetry is a reflection of her ignorance about objects and properties before the experiment. Given the lack of prior information, she has no reason to treat any two objects differently. This interpretation of symmetry is also known as the Laplacian Principle of Insufficient Reason, or Keynesian Principle of Indifference (Keynes (1921), Jaynes (1988), Savage (1972) chapter 3.7 and 4.5,

Kreps (1988) chapter 11, Gilboa, Postlewaite, and Schmeidler (2004)). Alternatively, there are many situations in which the symmetry assumption makes sense from the objective point of view. In particular, suppose that names of objects and properties are assigned to the objects randomly, without any attention to the state of the world. No relabelling should affect the observed frequency of outcomes. From the perspective of the DM, this looks *as if* the distribution of the states of the world were invariant with respect to any relabelling. I present examples of such situations.

Section 4 constructs a learning algorithm that has the features of the psychological model of categorization described above: the DM assigns objects into a few large categories and makes category-based predictions. The assignments into categories minimizes the inner-category entropy and maximizes the informational content of the categories. I compare two types of DMs. The Bayesian type knows the distribution ω and uses Bayes formula to make her predictions. The non-Bayesian type does not know ω and uses the categorization algorithm. In Section 4.4, I show that the non-Bayesian type makes asymptotically the same predictions as the Bayesian one, *no matter what the symmetric distribution* ω. In other words, the categorization algorithm achieves the quality of Bayesian prediction *uniformly* across all symmetric distributions. This leads to the first argument of this paper:

> **Argument 1.** Categorization is an optimal tool of prediction in symmetric environments.

There is an important statistical reason why the size of the categories increases with the number of incoming observations. Having fewer and larger categories helps with the problem of *overfitting*. Recall that the statistical literature warns against using high-dimensional models to "explain" limited observations, the reason being that we risk losing all the predictive power as a price for fitting the past data precisely. The use of fewer and larger categories alleviates this problem: classification of an object into the correct category is easier, given that there are many objects to compare it with; also, predictions in noisy categories are more precise, if based on a larger number of observations.

The categorization algorithm is not the only optimal solution to the prediction problem. In Section 4.5, I discuss a model in which the DM uses predictions to make decisions and to obtain payoffs. This allows the prediction problem to be cast as an exercise in payoff maximization under uncertainty. I demonstrate that all optimal solutions asymptotically lead to the same behavior. Hence,

> **Argument 2.** Any optimal behavior in symmetric environments is in the long-run behaviorally indistinguishable from categorization.

The above two arguments are concerned with the long-run optimality of categorization. In particular, it is optimal in the long-run to apply the Principle of Similarity and to predict

that objects with similar properties observed in the past are going to have similar properties in the future. Can one make any statements about the short-run behavior? The question is clearly relevant, as the various properties of similarity-based reasoning listed in the psychological literature (Rips (1975), Osherson, Wilkie, Smith, Lopez, and Shafir (1990)) come from experiments in which subjects have access to only very limited data samples. To address this question, I consider a Bayesian DM whose beliefs are symmetric. In Section 5, I describe the qualitative implications of prior symmetry and Bayesian rationality for predictions. The result is that

> **Argument 3.** The qualitative properties of Bayesian updating in symmetric environments coincide with experimental observations about similarity-based reasoning.

It needs to be emphasized that the symmetry assumption is motivated by the Principle of Insufficient Reason, not by any similarity considerations. This makes the coincidence between theoretical analysis and empirical observations somehow unexpected. In my opinion, the coincidence between theory and empirics is indirect evidence that the Principle of Insufficient Reason is strongly embedded in human reasoning.

There are three insights from the results of this paper. First, the fact that humans categorize is itself *not* evidence for bounded rationality. It is often a temptation to denounce heuristics as irrational and attribute using them to a mishandling of available information. On the contrary, categorization arises as an optimal statistical procedure. Second, categorization, if done optimally, does *not* lead to a persistent bias. The fully rational DM dynamically and endogenously adapts her categories to observations. There might be only a temporary bias, which is a consequence of insufficient data. Finally, it seems plausible that, through evolution, Nature equipped us with an optimal tool for making predictions in the face of uncertainty. The above results imply that if the natural environments are symmetric, then any such a tool should look like categorization.

2. Related literature

The literature on heuristics is divided between two strands. The first strand is concerned with the biases resulting from the use of heuristics. The most influential paper in this literature is Tversky and Kahneman (1974). Tversky (1977) presents a model of similarity-based reasoning. More specifically, biases of categorization have been studied in recent papers Mullainathan (2002), Lam (2000), Jackson and Fryer Jr (2005). Azrieli and Lehrer (2007) developed an axiomatic characterization of categorization. Jehiel (2005) (see also Jehiel and Samet (2007)) analyzes the consequences of categorization in a strategic context using a notion of analogy-basedexpectation equilibrium. The second strand argues that

heuristics are efficient ("fast and frugal") tools for processing information. For an example, see Gigerenzer and Todd (1999). The current paper belongs to the second strand of the literature.

The psychological literature mentions two major functions of categorization Smith (1995). First, as in this paper, categorization is a tool of inductive inference. Second, categorization serves as a device to code experience without being too demanding on our memory. It is probably the second role of the categorization that is more connected to bounded rationality.

I. Gilboa and D. Schmeidler's theory of case-based reasoning is an axiomatic approach to questions that are similar to the ones I ask here (Gilboa and Schmeidler (1995), Gilboa and Schmeidler (2000), Gilboa and Schmeidler (2001), Gilboa and Schmeidler (2002), Billot, Gilboa, Samet, and Schmeidler (2005)). These papers identify axioms under which the DM's behavior looks as if her prediction were guided by similarity between objects. The most important of these is the combination axiom, which says that if two different databases lead to the same prediction, their union should also lead to the same prediction. The combination axiom is controversial as many prediction rules (including the Bayes formula) do not satisfy it (see Gilboa and Schmeidler (1995) for a discussion). In particular, neither the categorization algorithm nor the Bayesian prediction of my model satisfies the combination axiom.[1] My paper contributes to the literature that the combination axiom is not necessary for similarity-based reasoning.

Collaborative filtering is a problem of predicting multiple properties of multiple objects from partial information about these and other objects and these and other properties Segaran (2007). One of the major applications of collaborative filtering methods are recommender systems. There are multiple customers (objects) with tastes over multiple products (properties). Each customer knows his tastes over certain set of products. The recommender system invites the customers to rate products they know and uses the ratings to predict the unknown tastes over the remaining products. For example, Netflix.com uses movie ratings of watched movies to make prediction about movies that has not been watched (Amazon.com and iTunes predict tastes over, respectively, books and music). All recommender system rely more or less directly on the Principle of Similarity. For example, in the earliest recommender systems, two customers were declared similar, if they were observed to have similar tastes over the same products and the system predicted that similar customers will have similar tastes over unobserved products: If one of the customer reveals that she likes the new product, the other customer will be predicted to like the new product as well. One of the major problems with this method is that in order to verify whether two customers are

[1] Gilboa, Lieberman, and Schmeidler (2005) argue that, as reasonable as it seems, the combination axiom should not hold when the DM "uses both inductive and deductive reasoning" at the same time.

similar, their tastes over the same set of products must be observed. In reality, the data available for the recommender system is often very sparse and it is rare for two randomly chosen customers to know their tastes over the same set of movies. For example, in the publicly available Netflix database, there are 100 million ratings of 480,000 customers over nearly 18,000 movies, and, an average customer rated slightly more than 1% of all available movies. In order to address this and related issues, other similarity-based algorithms were developed . For example, factor methods estimate model:

$$\theta = F_1 F_2' + E, \tag{2.1}$$

where θ is the $n_1 \times n_2$ matrix of tastes of n_1 customer over n_2 products, F_1 is a $n_1 \times k_1$ matrix of k_1 factor loadings for each customer, F_2 is a $n_2 \times k_2$ matrix of product factor loadings, $k_i << n_i$, and E is a matrix of i.i.d. noise terms (Canny (2002), Marlin and Zemel (2004)). Model (2.1) divides customer and products into categories of customers and products with the same factor loadings. The ratings of customers over products in the same (or similar categories) are equal up to i.i.d. noise terms. To the best of our knowledge, ours is the first paper that shows the asymptotic optimality of similarity-based prediction algorithms.[2]

3. Prior symmetry and principle of similarity

Let $X = X^1 \times X^2$ be a set of *instances* (inputs, independent variables, decision problems), where each instance (x^1, x^2) is a pair of two features x^1 and x^2. Assume that sets of features X^i are infinitely countable. Let $\{0,1\}$ be a set of *outcomes* (outputs, dependent variables, solutions). A *state of the world* is an assignment of an outcome to every instance, $\theta : X \to \{0,1\}$. Let $\Theta = \{0,1\}^X$ be the space of states of the world; Θ is a compact space in product topology, and it is a measurable space with Borel σ-field. A state of the world θ is chosen from distribution $\omega \in \Delta\Theta$. Consider the following examples.

Example 1 (Objects and Properties)**.** *Let $X^1 = O$ be a space of objects and $X^2 = P$ be a space of properties. Interpret instance $(o,p) \in O \times P$ as a query "Does object o have property p?" with an answer $\theta(o,p) \in \{0,1\}$. Say that objects o* and *o'* share property *p if $\theta(o,p) = \theta(o',p)$.*

Example 2 (Students and Grades)**.** *An undergraduate advisor helps students to predict grades. Each problem (instance, in my terminology) is described as a pair of two features*

[2]The Netflix algorithms described above are example of so-called *active filtering*, where the predictions are based on solicited customers' preferences. In *passive filtering*, the algorithm observes customer's choices and used these to make predictions. Kumar, Raghavan, Rajagopalan, and Tomkins (1998), Kleinberg and Sandler (2003), Kleinberg and Sandler (2004) are examples of papers that analyze the convergence properties of the passive filtering models. (I am grateful to the referee for pointing that literature to me.)

$x = (x^1, x^2) = (\textit{student, course})$. *An outcome* $\theta(x^1, x^2)$ *is equal to 1 if and only if student* x^1 *receives a good grade in course* x^2 *and* $\theta(x^1, x^2) = 0$ *if student* x^1 *receives a bad grade.*

Example 3 (Recommendation algorithms)**.** *Netflix.com helps costumers learn about their movie tastes using a recommender system: a customer rates movies, and Netflix uses the ratings of the customer and of other customers to predict a rating for the movies that the customer has yet not watched. (For a brief introduction into recommender systems, see Section 2.) Let* $X^1 = C$ *be a space of customers and* $X^2 = M$ *be a space of movies. Each instance* $(c, m) \in X$ *can be interpreted as a query "Is customer* c *interested in movie* m*?"*

3.1. **Symmetric distributions.** A *permutation of instances* π is a bijection of X onto itself. Denote the set of permutation of instances as Π. A *permutation* π^i *of dimension* i is a bijection of X^i onto itself. Denote the set of all permutations of dimension i as Π_i^F. A *permutation of features* is a mapping $\pi = \pi^1 \times \pi^2$, where $\pi^i \in \Pi_i^F$ for both i and

$$\left(\pi^1 \times \pi^2\right)\left(x^1, x^2\right) = \left(\pi^1\left(x^1\right), \pi^2\left(x^2\right)\right).$$

Thus, the permutation of features is a product of permutations of each dimension separately. Denote the set of all permutations of features with Π^F. Note that $\Pi^F \subseteq \Pi$, but $\Pi^F \neq \Pi$: not all permutations of instances are also permutations of features. (For example, suppose that $x \neq x'$; permutation $\pi_{x,x'} \in \Pi$, which exchanges instance x with x' and keeps all other instances constant, is not a permutation of features, $\pi_{x,x'} \notin \Pi^F$.)

For any permutation of instances π and any state of the world θ, define $\pi\theta \in \Theta$ as a state of the world, such that $(\pi\theta)(x) = \theta(\pi(x))$ for each instance x.

Definition 1. *Distribution* $\omega \in \Delta\Theta$ *is* symmetric (with respect to renaming features)*, if for any permutation of features* $\pi \in \Pi^F$*, for any measurable subset* $E \subseteq \Theta$*,*

$$\omega(\theta \in E) = \omega(\pi\theta \in E).$$

The symmetry condition generalizes exchangeability of de Finetti (1964) to two dimensions. It was introduced in Aldous (1981), Hoover (1982) (see also Kallenberg (2005)). The condition says that a priori all features are symmetric. In particular, no Bayesian DM considers any two features as a priori more similar to each other than to any other feature.

3.2. **Interpretation and examples.** From the subjective point of view, symmetry is a restatement of the Laplacian Principle of Insufficient Reason. It should be satisfied by any beliefs of the DM in a hypothetical state of perfect ignorance. In such a state, a Bayesian

DM hasn't yet observed any instances and outcomes, or anything that might be correlated with the state of the world. She has no reason to treat any two features differently a priori.

From the objective point of view, symmetry should not be interpreted as an assumption about the distribution from which Nature draws the state of the world (which would be quite restrictive) but about the DM's perception of it (which is not so restrictive). Imagine that (a) there is an objective state of the world, and (b) Nature randomly and uniformly mixes features before letting the DM observe outcomes. This leads to two labels of features: "original" and "perceived." If the mixing is truly uniform, then, from the point of view of the DM, the "perceived" feature x^i looks like the "original" x^i with the same probability as it looks like the "original" $x^{i'}$.

To see this argument more clearly, consider Example 2: Suppose that the registrar office randomly assigns ID numbers to students and courses. The DM knows the IDs but not the individual names of students or courses. If the assignment is completely random, then, from the point of view of the DM, it looks *as if* the state of the world is drawn from a symmetric distribution.

As an another example, consider the Netflix problem from Example 3. From the point of view of Netflix, no renaming of its 5 million customers should change the correlations among customers, movies and their preferences. Analogously, no renaming of their 65 000 movie titles should affect the distribution of the states of the world.[3]

Since a priori all features are exchangeable, the assumption precludes any possibility of non-empirical categorization. However, the assumption allows for a wide range of theories about correlations between outcomes of instances. These correlations may open a possibility of ex post categorization. Consider the following examples. In the first example, all correlations are eliminated. Two subsequent examples are more sophisticated.

Example 4 (Idiosyncratic preferences)**.** *In the Netflix example, suppose that each outcome* $\theta(x)$ *is chosen i.i.d. from uniform distribution on* $\{0,1\}$*, independently across instances.*

Example 5 (Bad and good movies)**.** *Suppose that each movie* $m \in X^2$ *is independently chosen to be good with probability* $p \in [0,1]$ *or bad with probability* $1-p$*. State of the world* θ *depends deterministically on the quality of movies: for any instance* $(c,m) \in X$*, let* $\theta(c,m) = 1$ *if movie* m *is good; otherwise, let* $\theta(c,m) = 0$*.*

[3]Recently, Netflix announced a public competition for a recommendation algorithm that improves on its own (see www.netflixprize.com). A database of 100 million customer-movie rankings is available for any contestant. In order to protect the confidentionality of ratings, Netflix replaced the customers' and movies' names by randomly drawn IDs. In other words, from the perspective of the contestant, the Netflix database looks as if it were drawn from an invariant distribution.

Example 6 (Two types of movies and customers)**.** *There are two types of customers, Men and Women and two types of movies, Action and Romance. Each customer is chosen independently to be Man or Woman with probability* $\frac{1}{2}$*; similarly, each movie is chosen to be Action or Romance with equal probability. State of the world* θ *depends on the types of customers and movies:*

$$\theta(c,m) = \begin{cases} 1, & \text{if } c \text{ is Man and } m \text{ is Action or } c \text{ is Woman and } m \text{ is Romance,} \\ 0, & \text{otherwise.} \end{cases}$$

In the examples, customers and movies are divided into types (categories). In the first example, there is only one category for customers and movies; in the second, there is one category for customers and two categories for movies; in the last, there are two categories for customers and movies. The outcomes depend on the category assignment either probabilistically (as in the first example) or deterministically (as in the two subsequent examples.) Appendix B.1 contains the Representation Theorem for symmetric distribution. The Theorem shows that any symmetric distribution is a mixture of distributions generated as in the examples, but with, possibly, infinitely many categories.

3.3. **Principle of Similarity.** Recall that the Principle of Similarity says that if two objects were observed to have similar properties, their unobserved properties should be expected to be similar. Without any further analysis, it is unclear what the Principle has to do with symmetric distribution. Nevertheless, the connection can be illustrated with a simple result. Consider Example 1. Suppose that the Bayesian DM with symmetric beliefs ω observes properties $p \in P'$ of two objects $o_1, o_2 \in O$, where P' is finite set. The DM wonders whether these objects share an unobserved property $p^* \notin P'$. The next Proposition says that the probability of such an event increases in the number of shared properties $p \in P'$.

Proposition 1. *For any symmetric* ω*, any two sets* $P_1, P_2 \subseteq P'$*, if* $|P_1| \leq |P_2|$*, then for all* $p^* \notin P_1 \cup P_2$

$$\begin{aligned} &\omega\left(\theta(o_1,p^*) = \theta(o_2,p^*) \,|\, \theta(o_1,p) = \theta(o_2,p) \;\; \text{if } p \in P_1\right) \\ &\leq \omega\left(\theta(o_1,p^*) = \theta(o_2,p^*) \,|\, \theta(o_1,p) = \theta(o_2,p) \;\; \text{if } p \in P_2\right). \end{aligned}$$

Proof. Consider a random variable $s : P \to \{0,1\}$ defined as a function of the state of the world: for any property p, $s(p) = 1$ if and only if objects o_1 and o_2 share property p, i.e. $\theta(o_1,p) = \theta(o_2,p)$. Let $\varpi \in \Delta\{0,1\}^P$ be the distribution of variable s induced by symmetric distribution ω. The symmetry of ω implies that distribution ϖ is invariant with respect to permutation of properties: $\varpi(s \in E) = \varpi\left(\pi^P s \in E\right)$ for any bijection $\pi^P : P \to P$ and any measurable $E \subseteq \{0,1\}^P$. By de Finetti's Theorem, ϖ can be interpreted as a distribution of infinitely many Bernoulli draws indexed with $p \in P$, with a parameter that is stochastically

drawn once from some $\mu \in \Delta[0,1]$. It is a simple consequence of the representation that the conditional probability of $s(p^*) = 1$ increases with the number of 1s observed so far,

$$\begin{aligned}&\omega(s(p^*) = s(p^*) \,|\, s(p) = s(p) \text{ iff } p \in P_1)\\ &\leq \omega(s(p^*) = s(p^*) \,|\, s(p) = s(p) \text{ iff } p \in P_2).\end{aligned}$$

This yields the Proposition. □

The Proposition provides the main intuition for the connection between symmetry and similarity-based reasoning. All the subsequent results can be seen as generalizations of this intuition.

3.4. **Examples without symmetry.** In many cases, the symmetry assumption is too strong and the DM has an a priori information that can be used to distinguish individual objects. For example, Netflix.com may distinguish movies by their genres, the items sold on the Amazon.com website can be exogenously categorized into Books, Movies, or Computers, and so on ... . The results of this paper go unchanged as long as the prior information is not too detailed. Precisely, say that distribution $\omega \in \Delta\Theta$ is *finitely symmetric* if there are finite partitions $X^i = \bigcup_{k \leq K} X^{i,(k)}$, such that for each $k, l \leq K$, the marginal distribution

$$\text{marg}_{\{0,1\}^{X^{1,(k)} \times X^{1,(l)}}} \omega$$

is symmetric. Such a distribution allows for limited a priori categorization of objects and properties where the number of prior categories is bounded by the size of the partition. Up to minor modifications of the proof, the main results of this paper (Theorem 1 and Corollary 1) hold.

There are examples, in which the prior information is too detailed. For an extreme example, suppose that X^1 is the set of people, X^2 is the set of natural numbers, and $\theta(x^1, x^2) = 1$ if person x^1 is alive at the age x^2. The prior information does not distinguish between people X^1; however, the prior information imposes non-symmetric structure on X^2. Because elements of X^2 are not exchangeable, the categorization algorithm of this paper won't apply.

4. Categorization

4.1. **Model of learning.** We discuss categorization as a DM's strategy in the following model of learning. A sequence of instances $\bar{x} = x_1, x_2, ... \in X^\infty$ is called an *instance process*. To avoid trivial cases, I assume that the DM never observes the same instance twice, $x_s \neq x_t$ for $s \neq t$. Each period, the DM observes an instance x_t, makes a prediction and subsequently observes outcome $y_t = \theta(x_t)$. *Learning rule* $l : \bigcup_t (X \times \{0,1\})^{t-1} \times X \to \Delta\{0,1\}$ is a complete description of the predictive behavior of the DM. The predicted probability that the

outcome of x_t is equal to y is denoted as $l\left(\{x_s, y_s\}_{s<t}, x_t\right)(y)$. In particular, each distribution ω induces a Bayesian learning rule

$$l_\omega\left((x_s, y_s)_{s<t}, x_t\right)(y) := \omega\left(\theta(x_t) = y \middle| \{x_s, y_s\}_{s<t}\right).$$

A *database* d is any finite subset of observations $d \subseteq X \times \{0,1\}$. The size of database d is denoted with $|d|$. For any database d, let

$$d^i = \left|\left\{x^i \in X^i : \left(x^i, x^j, y\right) \in d \text{ for some } \left(x^i, x^j\right) \in X \text{ and } y \in Y\right\}\right|$$

denote the number of distinct features i in database d. In particular, given an instance process $\bar{x}$ and state of the world θ, the period t database of past observations is defined as $d_t = \{(x_s, \theta(x_s))\}_{s<t}$ and $|d_t| = t$. If the value of the learning rule does not depend on the order of past observations,[4] write $l(d, x)$ for database d and instance x. For example, Bayesian learning rule l_ω does not depend on the order of observations.

Definition 2. *Instance process $\bar{x}$ satisfies* sufficient data condition *if*

$$\frac{t}{d_t^1 + d_t^2} \to \infty.$$

The sufficient data condition implies that the number of observations grows quicker than the number of distinct features in the database of past observations. In the Netflix example (Example 3), this means that the number of observations per customer and per movie increases to infinity.

4.2. **Categorization algorithm.** Next, I construct two learning rules. In both rules, the DM divides instances into finitely many categories. The number of categories is fixed in the first learning rule, and it increases with the number of observations in the second one. The rules share stylized characteristics with the categorizing behavior discussed in the psychology literature: (a) categorization is endogenous and dynamic, (b) the number of categories is small compared to the number of objects, and (c) the prediction is category-based: all objects in the same category are predicted to share similar properties.

Any categorization process must solve two difficulties: how to allocate instances into categories and how to find a prediction of an outcome conditional on the category. Both difficulties are addressed simultaneously. There are k possible "bins" for features i and each feature is assigned into only one bin. If features x^1 and x^2 are assigned to categories k^1 and k^2, respectively, then the outcome $\theta(x^1, x^2)$ is predicted to be 1 with probability $\rho(k^1, k^2) \in [0,1]$.

[4]More precisely, the value of the learning rule l does not depend on the order of past observations, if for each t, each sequence $((x_1, y_1), ..., (x_t, y_t))$, each x_{t+1}, each bijection $\pi : \{1, ..., t\} \to \{1, ..., t\}$,

$$l((x_1, y_1), ..., (x_t, y_t), x_{t+1}) = l\left(\left(x_{\pi(1)}, y_{\pi(1)}\right), ..., \left(x_{\pi(t)}, y_{\pi(t)}\right), x_{t+1}\right)$$

Initially, the DM is uncertain which assignment into bins and which prediction function ρ are the best, i.e. the most helpful in facilitating predictions. She acts as a Bayesian: she starts with an uniform prior over all possible assignments and functions ρ. When new information comes, she updates her prior through Bayes formula.

Formally, let *k-(category) assignment* of feature i be a map $c^i : X^i \to \{1, ..., k\}$. I refer to $c^i(x^i)$ as a category of feature x^i. Let $\mathcal{C}_i^k = \{1, ..., k\}^{X^i}$ be a set of k-assignments of feature i and let $\mathcal{C}^k = \mathcal{C}_1^k \times \mathcal{C}_2^k$ be a set of k-assignments. For any k-assignment $c \in \mathcal{C}^k$, any instance $x = (x^1, x^2) \in X$, write

$$c(x) = \left(c^1\left(x^1\right), c^2\left(x^2\right)\right) \in \{1, ..., k\}^2$$

and call $c(x)$ a category of instance x with respect to assignment c. Hence, k-assignment divides instances into k^2 categories.

A *(category-based) prediction* is a function $p : \{1, ..., k\}^2 \to \Delta\{0, 1\}$ with the following interpretation: if the DM decides to assign instance x to category $\mathbf{k} \in \{1, ..., k\}^2$, then she predicts that the outcome of x is equal to y with probability $p(\mathbf{k})(y)$. Define space of prediction functions as

$$\mathcal{R}^k := \left\{\rho : \{1, ..., k\}^2 \to \Delta\{0, 1\}\right\} = [0, 1]^{k^2}.$$

A pair of an assignment and a prediction function is called a *theory.* Let

$$\mathcal{T}^k = \mathcal{C}^k \times \mathcal{R}^k$$

be a space of theories.

The DM starts with a prior beliefs over theories. Let $\Psi_i^k \in \Delta\mathcal{C}_i^{k_i}$ be the "uniform" measure over assignments of i. It is formally defined as a measure such that for any feature $x^i \in X^i$ category $c^i(x^i)$ is drawn independently and uniformly from set $\{1, ..., k\}$. Let

$$\Psi_C^k = \Psi_1^k \otimes \Psi_2^k \in \Delta\mathcal{C}^k$$

be the independent product of two measures. Distribution Ψ_C^k is the uniform measure over a space of assignments $\mathcal{C}^k$. Let Ψ_R be the Lebesgue measure on the space of prediction functions $\mathcal{R}^k$. Let

$$\Psi^k = \Psi_C^k \otimes \Psi_R^k \in \Delta\mathcal{T}^k$$

be the independent product of measures Ψ_C^k and Ψ_R^k. Distribution Ψ^k is the uniform measure over an infinite dimensional space of theories and it is treated as the prior beliefs.

Consider a distribution ω^k over outcomes $\theta(.)$ defined in the following way: First, Nature chooses a theory (c, ρ) from distribution Ψ^k. Then, for each $x \in X$, each outcome $\theta(x)$ is drawn independently from each other from distribution $\rho(c(x))$. It is easy to check that

distribution ω^k is symmetric. Define learning rule l^k as the prediction that would be made by a Bayesian DM with "beliefs" ω^k :

$$l^k = l_{\omega^k}.$$

I refer to learning rule l^k as a *k-categorization algorithm*.

So far, I have assumed that the number of categories k remains constant. More generally, the DM may want to vary k with the number of data. She needs to be aware of two effects of increasing k. On one hand, a higher k allows for a better fit with the sample data allowing possibly for sharper predictions. On the other hand, too high a k may lead to the *overfitting* problem: the more categories she has, the more difficult it is to use finite data to choose the right category assignment and the right prediction function. (For example, if in each period t, the DM uses t categories, clearly his predictions will not converge to the Bayesian prediction, except for possibly some trivial cases.)

In the solution that I propose, the DM acts as if he starts with a prior belief over the number of categories k. Fix a sequence of strictly positive real numbers $\alpha_k > 0$ such that $\sum_k \alpha_k = 1$.[5] Let

$$\omega_C = \sum_k \alpha_k \omega^k$$

be the mixture of probability distributions with weights k. Define *the adaptive categorization algorithm* as the Bayesian learning rule corresponding to distribution $l_C = l_{\omega_C}$. The interpretation is that, initially, the beliefs of the DM are concentrated over the theories with small number of categories. When the database increases, the DM may notice that low k do not explain the data very well; he will update his beliefs to put more weight on larger k. [6]

4.3. **Categorization as entropy minimization.** It is helpful to reinterpret the categorization algorithm l^k as entropy minimization. For each database d and category assignment c, define the number of instances assigned to category (k^1, k^2) and the frequency of outcome 1 among these instances as

$$n\left(k^1, k^2 | c, d\right) = \#\left\{\left(x^1, x^2, y\right) \in d : c\left(x^1, x^2\right) = \left(k^1, k^2\right), y \in \{0, 1\}\right\},$$
$$\phi\left(k^1, k^2 | c, d\right) := \frac{\#\left\{(x^1, x^2, 1) \in d : c\left(x^1, x^2\right) = (k^1, k^2)\right\}}{n\left(k^1, k^2 | c, d\right)},$$

[5] More precisely, it is enough that infinitely many weights are strictly positive.

[6] I am grateful to the referee for suggesting learning rule l_C as well as the strategy of the proof of Theorem 1. A previous version of the paper analyzed a non-Bayesian learning rule l_C

if $n\left(k^1,k^2|c,d\right)\neq 0$ and $\frac{1}{2}$ otherwise. Define the *entropy* of assignment c as

$$E\left(c|d\right)=-\frac{1}{|d|}\sum_{(k^1,k^2)\in\{1,...,k\}^2} n\left(k^1,k^2|c,d\right)h\left(\phi\left(k^1,k^2|c,d\right)\right),$$

where $h\left(.\right)$ is the entropy function

$$h\left(\phi\right)=\phi\log\phi+\left(1-\phi\right)\log\left(1-\phi\right).$$

Entropy $E\left(c|d\right)$ measures the informational content of assignment c in database d. In particular, if the entropy is close to 0, then, for most categories, $\phi\left(k^1,k^2\right)$ is very close to 0 or very close to 1, i.e. predictions inside categories are almost deterministic. On the other hand, if the entropy is close to $\log 2$ (which is the maximal possible value), then $\phi\left(k^1,k^2\right)$ is close to $\frac{1}{2}$ and categories are quite useless in facilitating prediction. Define the minimal value of entropy in database d as

$$E_{\min}\left(d\right)=\min_{c\in\mathcal{C}^k}E\left(c|d\right).$$

The next Proposition says that, if the sufficient data condition is satisfied, then, asymptotically, the categorization algorithm puts probability close to 1 to the set of assignments with entropy close to the minimal entropy.

Proposition 2. *Suppose that instance process $\bar{x}$ satisfies the sufficient data condition. Then, for any $\varepsilon>0$,*

$$\lim_{t\to\infty}\int_{\{c:E(c|d_t)\leq E_{\min}(d_t)+\varepsilon\}\times\mathcal{R}^k}\psi^k\left(c,\rho|d_t\right)d\Psi^k\left(c,\rho\right)=1.$$

The (standard) proof is contained in Appendix A. Jackson and Fryer Jr (2005) is also concerned with the issue of dividing objects into categories. The authors postulate a simple heuristic: find a categorization assignment that minimizes variance inside categories. Since entropy minimization is not the same as minimization of inner-categroy variance, Jackson and Fryer Jr (2005)'s algorithm is not going to satisfy the optimality results.

4.4. **Optimality of categorization.** This section shows that categorization is an optimal solution to the prediction problem. Consider two types of DMs. A Bayesian DM knows which symmetric distribution ω generates the state of the world. The best prediction she can make is to use Bayes learning rule l_ω. A non-Bayesian DM does not know the true distribution ω, but nonetheless believes that ω is symmetric.

The next Proposition says that if k is sufficiently high, then the non-Bayesian DM who uses k-categorization algorithm l^k makes asymptotically similar predictions to those made by the Bayesian DM. It is convenient to evaluate the difference between two measures $p,q\in\Delta\left\{0,1\right\}$

by their L^2-distance $\|p-q\| = \left(\sum_{y\in\{0,1\}} \left(p\left(y\right)-q\left(y\right)\right)^2\right)^{1/2}$. Let E_ω denote the expectation is taken with respect to the distribution over past databases induced by ω (and instance process $\bar{x}$).

Theorem 1. *Suppose that process $\bar{x}$ satisfies the sufficient data condition. For any symmetric ω,*

$$\lim_{k\to\infty}\lim_{t\to\infty}\frac{1}{t}E\sum\nolimits_{s<t}\left\|l^k\left(d_s,x_s\right)-l_\omega\left(d_s,x_s\right)\right\|=0,$$
$$\lim_{t\to\infty}\frac{1}{t}E\sum\nolimits_{s<t}\left\|l_C\left(d_s,x_s\right)-l_\omega\left(d_s,x_s\right)\right\|=0. \tag{4.1}$$

By the first part of the Theorem, any symmetric ω can be approximated by a model with a finite number of categories. For a k high enough, additional categories do not substantially increase the predictive power of the categorization algorithm. Note that the number of categories k needed depends on the distribution ω. The second part says that the adaptive categorization l_C makes on average the same predictions as the Bayesian DM *uniformly* across all symmetric beliefs ω.

To interpret the Theorem, consider first the subjective point of view on ω. Any Bayesian DM with symmetric beliefs expects to predict *as if*, approximately and asymptotically, she was using the categorization algorithm. Therefore, any Bayesian DM is indifferent between Bayesian updating and categorization.

From the objective viewpoint, Nature draws the state of the world from symmetric distribution ω. The DM may understand that the distribution is symmetric, even if she does not know ω. The categorization algorithm guarantees payoffs as good as if the DM knew the true ω. This is very good news for any ambiguity-averse decision maker.

Note that the Theorem does not guarantee that the DM will predict all outcomes correctly but only as good as the Bayesian. In Example 4, all outcomes are i.i.d. equal to 1 with probability $\frac{1}{2}$. This is the prediction of the Bayesian DM no matter how much past data she observes. By the Theorem, this is also the asymptotic prediction of the categorization algorithm l_C. On the other hand, if the sufficient data condition is satisfied, then, asymptotically, the categorization algorithm predicts correctly almost all outcomes in Examples 5 and 6.

The statement of the Theorem is related to the literature on the Bayesian merging of opinions (Blackwell and Dubins (1962), Lehrer and Smorodinsky (1996), Lehrer and Smorodinsky (2000); also Jackson, Kalai, and Smorodinsky (1999)). The proof uses the Representation Theorem from Appendix B.1 as well as the techniques developed in Lehrer and Smorodinsky (1996) and Lehrer and Smorodinsky (2000). Appendix C contains the details as well as the discussion of the connection to the earlier work.

4.5. **Uniqueness of optimal solution.** Theorem 1 shows that there is a learning rule that guarantees uniformly good predictions. Moreover, this rule, by construction, can be interpreted as a categorization. The Theorem does not guarantee that such a rule is unique, and, in fact, there are infinitely many such learning rules. This is a consequence of the chosen criterion of optimality: any learning rule that behaves differently than l_C in the first t periods and then follows the same predictions as l_C satisfies the formula (4.1).

Let A be a compact and normed space of actions and $u : A \times \{0,1\} \to R$ be a continuous utility function, such that the solution to the optimization problem

$$\max_a (1-p)\, u(a,0) + pu(a,1)$$

exists, and it is unique and continuous in p.[7] Denote the solution as $a_{\max}(p)$. Let $a : \bigcup_t (X \times \{0,1\})^{t-1} \times X \to R$ designate a *behavioral rule*. Let

$$u\left(a\left(\left(x_s, \theta(x_s)\right)_{s<t}, x_t\right), \theta(x_t)\right)$$

be a payoff in period t from behavioral rule a in the state of the world θ. For any distribution ω and any instance process $\bar{x}$, let

$$U(a;\omega,\bar{x}) := \lim_{t\to\infty} \inf E \frac{1}{t} \sum_{s<t} u\left(a\left(\left(x_s, \theta(x_s)\right)_{s<t}, x_t\right), \theta(x_t)\right)$$

denote a long-run expected quality of behavioral rule a, where the expectation is taken with respect to the distribution over databases of past observations induced by ω and instance process $\bar{x}$. Let

$$a^l := a_{\max} \circ l$$

denote the behavioral rule induced by learning rule l.

Definition 3. *Behavior a is* uniformly optimal *if $U(a;\omega,\bar{x}) \geq U(a';\omega,\bar{x})$ for any symmetric distribution ω and for any other behavior a'.*

The definition of uniformly optimal behavior is very strong. It requires the behavior to be (weakly) better than any other behavior for any other distribution over states of the world. Any uniformly optimal behavior is robust to misspecification of prior beliefs. The notion of robustness is stronger if the behavior was simply optimal with respect to the minimax preferences of Gilboa and Schmeidler (1989). In the latter, the DM cares only about the worst-case payoff. Here, the DM achieves the optimal payoff given any distribution ω.

[7] For example, suppose that $A = [0,1]$ and $u(a,y) = ay - \frac{1}{2}a^2$.

Premise	Robins use serotonin as a neurotransmitter.
	Bluejays use serotonin as a neurotransmitter.
Conclusion	Sparrows use serotonin as a neurotransmitter.

Premise	Robins use serotonin as a neurotransmitter.
	Bluejays use serotonin as a neurotransmitter.
Conclusion	Geese use serotonin as a neurotransmitter.

TABLE 2. Premise-conclusion similarity

Corollary 1. *Categorization behavior a^{lC} is uniformly optimal. For any process $\bar{x}$, for any uniformly optimal behavior a,*

$$\limsup_{t\to\infty} \frac{1}{t} E \sum\nolimits_{s<t} \left\| a^{lC}\left(d_s, x_s\right) - a\left(d_s, x_s\right)\right\| = 0.$$

The Corollary says that all uniformly optimal behavioral rules are asymptotically equal and, in particular, all of them are equal to the categorization algorithm. The idea is very simple. The best prediction possible given any symmetric ω is the Bayesian prediction. Because categorization makes the Bayesian prediction asymptotically, any uniformly optimal behavior must do the same. Therefore, any two uniformly optimal behaviors are asymptotically equal. In particular, any uniformly optimal behavior is asymptotically indistinguishable from categorization.

Proof. Let $a_\omega := a_{\max} \circ l_\omega$ denote Bayesian behavioral rule. By standard arguments, for any instance process $\bar{x}$, any symmetric ω and any behavioral rule a,

$$U\left(a; \omega, \bar{x}\right) \leq U\left(a_\omega; \omega, \bar{x}\right)$$

with strict inequality if

$$\limsup_{t\to\infty} \frac{1}{t} E \sum\nolimits_{s<t} \left\| a_\omega\left(d_s, x_s\right) - a\left(d_s, x_s\right)\right\| > 0.$$

The inequality, together with Theorem 1, implies the thesis of the Corollary. □

5. SIMILARITY-BASED REASONING

Osherson, Wilkie, Smith, Lopez, and Shafir (1990) performs a series of laboratory experiments that aim to systematize features of similarity-based reasoning and categorization. In each of individual experiments, they present subjects with a pair of inductive arguments. Each argument consists of a premise followed by a conclusion. Subjects are asked to choose the more credible argument in a pair. An example of such a comparison is presented in Table 2.

Premise	Robins use serotonin as a neurotransmitter.
	Bluejays use serotonin as a neurotransmitter.
	Robins, bluejays, and sparrows are small.
	Geese are large.
Conclusion	Sparrows use serotonin as a neurotransmitter.

Premise	Robins use serotonin as a neurotransmitter.
	Bluejays use serotonin as a neurotransmitter.
	Robins, bluejays, and sparrows are small.
	Geese are large.
Conclusion	Geese use serotonin as a neurotransmitter.

TABLE 3. Premise-conclusion similarity

There are some features of the experiment design that are worth emphasizing. Each statement in premise or conclusion refers to object of natural kind (most often animals) having or not certain *blank* property. The blank properties are chosen so to minimize the prior knowledge of these properties that the subjects (students of psychology) bring to the experiment. Specifically, subjects are expected to realize that all properties in question are anatomical, but they are not supposed to know much more. For example, property p ='use serotonin as a neurotransmitter' from Table 2 is blank. (It is rare for non-professionals to be able to differentiate prior probabilities of "using" and "not using serotonin as a neurotransmitter.") By definition, subjects have "no reason to believe" that various blank properties can be distinguished a priori. In the language of this paper, the beliefs about blank properties are symmetric with respect to their renaming.

On the other hand, subjects are expected to have prior knowledge of objects, and, specifically, share notion of similarity between objects. For example, 'robins' and 'bluejays' are considered as more similar than 'robins' and 'geese.' It seems natural to assume that prior knowledge comes from prior observation that 'robins' and 'bluejays' share some properties, like 'being small', with each other but not with 'geese'. In particular, imagine a non-English speaking subject who has never heard about 'robins', 'bluejays' or 'geese', and who is asked to evaluate inductive arguments in Table 3. Such subject should reason in the same way as the English-speaker presented with Table 2. The non-English speaker's beliefs are symmetric with respect to *both* objects and properties.

We interpret experiments as in Table 2 in the context of example 1. We assume that subjects have initially symmetric beliefs about *both* objects and properties. Subjects are asked about the probability whether an object has certain property conditionally on two pieces of

Premise	Bluejays require Vitamin K for the liver to function. Falcons require Vitamin K for the liver to function.
Conclusion	All animals require Vitamin K for the liver to function.

Premise	Bluejays require Vitamin K for the liver to function. Falcons require Vitamin K for the liver to function.
Conclusion	All birds require Vitamin K for the liver to function.

TABLE 4. Premise diversity

prior information: prior knowledge about similarity with respect to known properties that is brought to experiment, and information about a blank property given by the experimenter. Below, we discuss three examples of such experiment. In the first two cases, the theoretical predictions coincide to some degree with the laboratory observations.

5.1. **Premise-conclusion similarity.** Faced with comparison from Table 2, 59 out of 80 subjects choose the first argument as the more credible. The interpretation is that a category of sparrows is more similar to categories of robins and bluejays and the corresponding inductive argument seems more appropriate. As we argue above, it seems reasonable to assume that the subjects beliefs are symmetric, and to treat the reasoning in Table 2 as an application of the Principle of Similarity and Proposition 1.

Further, Osherson, Wilkie, Smith, Lopez, and Shafir (1990) argue that an inductive argument is more credible if the conclusion is more specific (and more similar to the premise.). Consider an example in Table 4. Here, 75 out of 80 subjects point to the second argument as more credible.

This experiment differs from the previous one because here, the conclusion concerns a class of rather than a specific object. Nevertheless, conclusion specificity is an application of the Principle of Similarity (as formulated as, possibly a variation of, Proposition 1).

5.2. **Premise diversity.** The next experiment indicates that an inductive argument is more credible if the premise is more diverse. Consider an example in Table 5. Here, 76 out of 80 subjects point to the first argument as more credible.[8]

The premise diversity is somehow surprising, given what has been said so far about similarity-based reasoning. Nevertheless, it has a consistent explanation. Given that rhinoceroses are known a priori to be similar to hippopotamuses, *it is not unexpected* that rhinoceroses and

[8]It is instructive to compare premise-diversity with an observation reported in Glazer and Rubinstein (2001) that subjects believe that the counterargument to some thesis is more credible if the object is more similar to the object in the original argument. It is difficult to explain Glazer and Rubinstein's phenomenon in a Bayesian setting and Glazer and Rubinstein (2001) propose a game-theoretic explanation.

Premise	Hippopotamuses have a higher sodium concentration in their blood than humans.
	Hamsters have a higher sodium concentration in their blood than humans.
Conclusion	All mammals have a higher sodium concentration in their blood than humans.

Premise	Hippopotamuses have a higher sodium concentration in their blood than humans.
	Rhinoceroses have a higher sodium concentration in their blood than humans.
Conclusion	All mammals have a higher sodium concentration in their blood than humans.

TABLE 5. Premise diversity

hippopotamuses have similar amounts of sodium. In particular, the fact about rhinoceroses does not add to what is already known from the analogous statement about hippopotamuses. On the other hand, the same statement about hamsters is more informative: it signals that "high sodium concentration" is shared by other mammals.

As above, consider Example 1. Take any three objects o, o_A, o_B, and any property p. A natural way to formalize the premise diversity is to ask how the probability that object o shares property p with objects o_A and o_B depends on the similarity of o_A and o_B: is it true that

$$\begin{aligned}&\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right)|\theta\left(o_A,p\right)=\theta\left(o_B,p\right)\text{ and }o_A,o_B\text{ are similar}\right)\\&\leq\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right)|\theta\left(o_A,p\right)=\theta\left(o_B,p\right)\text{ and }o_A,o_B\text{ are not similar}\right)?\end{aligned}\tag{5.1}$$

This inequality does not hold generally for all symmetric ω.[9] This is because the probabilities in (5.1) are conditioned not only on the information about the similarity/diversity of the premise, but also on the fact that two randomly chosen objects o_A and o_B are either similar or diverse. It is a one of the applications of the Principle of Similarity that the probability that o is similar to o_A (with respect to property p) increases (or decreases) if the objects in the population tend to be similar to each other (or different from each other).

To isolate the effect of premise similarity/diversity, consider a slightly more complicated question. Take five objects $o, o_A, o_B, o_C, o_D \in O$ and two properties $p, p' \in P$. For example, property p may correspond to 'higher sodium concentration in their blood than humans',

[9] As a simple example, consider a distribution ω as an equal mixture of two distributions ω_1 and $\omega_{1/2}$, where ω_1 chooses all outcomes to be equal to 1 and $\omega_{1/2}$ chooses i.i.d. outcomes of all instances to be equal to 0 or 1 with the same probability. Then, the information that objects o_A and o_B are not similar makes certain that outcomes are chosen from distribution $\omega_{1/2}$ and the right-hand side of (5.1) is equal to $\frac{1}{2}$. On the other hand, the left-hand side of (5.1) is strictly higher than $\frac{1}{2}$.

Mice have a lower body temperature than humans
Bats have a lower body temperature than humans

Bats have a lower body temperature than humans
Mice have a lower body temperature than humans

TABLE 6. Premise-conclusion asymmetry

and property p' may correspond to 'living in Africa'. Define events:

$$P^* = \{\theta(o_A, p') = \theta(o_B, p') \neq \theta(o_C, p') = \theta(o_D, p')\}$$
$$PS = \{\theta(o_A, p) = \theta(o_B, p)\} \cap P^*,$$
$$PD = \{\theta(o_A, p) = \theta(o_C, p)\} \cap P^*.$$

Event P^* defines two different pairs of similar objects $\{o_A, o_B\}$ and $\{o_C, o_D\}$. In event PS, two similar objects o_A and o_B share an additional property p. I refer to evidence about property p as a premise and to this event as a similar premise. In event PD, two different objects o_A and o_C share property p. I refer to that event as a diverse premise.

Proposition 3. *For any symmetric* ω,

$$\omega(\theta(o, p) = \theta(o_A, p) \,|PS) \leq \omega(\theta(o, p) = \theta(o_A, p) \,|PD).$$

The proof can be found in Appendix D. The Proposition compares the probability of the fact that objects o and o_A share property p conditional on similar or diverse premise: The probability that o and o_A share property p is higher when the premise is diverse.

5.3. **Premise-conclusion asymmetry.** Finally, I present an example of similarity-based reasoning that cannot be captured in the model of this paper. Consider the inductive arguments in Table 6. Osherson, Wilkie, Smith, Lopez, and Shafir (1990) report that a majority of students select the first argument as more credible.[10] Following Rips (1975), they argue that "mice" are more typical animals; hence it is more informative about properties shared by general category of similar animals. "Bats" are exotic, and it is not surprising that they have exotic properties.

No Bayesian model exhibits consistent asymmetry of this form. To see it, suppose that the DM has a probability distribution over whether "mice" and "bats" have property p:

[10]The results seem to be very weak. In the first round of the experiment, 41 out of 80 students pointed to the first argument and 39 pointed to the second. Only in the second round, when the subjects were explicitly instructed "Although the arguments may seem similar, there is always a difference in how much reason the facts of an argument give to believe its conclusions", did 40 out of 60 students select the first argument.

$\mu \in \Delta\left(\{0,1\}^{\{M,B\}}\right)$. Then the conditional probabilities conditional on the premise in the first and second argument are equal, respectively, to

$$\frac{\mu\left(y\left(M\right)=1, y\left(B\right)=1\right)}{\mu\left(y\left(M\right)=1\right)} \text{ and } \frac{\mu\left(y\left(M\right)=1, y\left(B\right)=1\right)}{\mu\left(y\left(B\right)=1\right)}.$$

In the benchmark case, the probabilities that "mice" and "bats" have property p should be equal, $\mu\left(y\left(M\right)=1\right) = \mu\left(y\left(B\right)=1\right)$. (One can think about it as an application of the Principle of Insufficient Reason: Notice that property p "have lower body temperature than humans" seems to be exchangeable with its logical negation $-p$ "have higher body temperature than humans".) But then both conditional probabilities must be equal.

6. Extensions

In this Section, I discuss possible extensions of the model. Details of some of the generalizations can be found in the previous version of the current paper Peski (2006).

Finite space of outcomes. So far, a state of the world has been defined as a mapping $\theta : X \to Y$, where the space of outcomes $Y = \{0,1\}$ is binary. As a simple extension, consider any finite space Y. The results do not change and the proofs change in a predictable way. For example, the prediction function in Section 4 should be redefined as a mapping $\rho : \{1,...,k\}^2 \to \Delta Y$.

Stochastic instance process. Suppose that instances are drawn from a stochastic distribution $\mu_X \in \Delta X^\infty$. Say that the sufficient data condition is satisfied for μ_X if it is satisfied almost surely for each of the realizations. If one assumes that the path of instances $x_1, x_2, ...$, is drawn independently from the realization of the state of the world θ, then all the results hold μ_X-surely.

The independence of instance process and the state of the world achieves the following goal. Consider a scientist who designs experiments, i.e. chooses the instance process, and whose goal is not to predict well, but to find interesting observations. Such a scientist will be interested mostly in outcomes of instances that are difficult to predict. This is because such outcomes are probably most interesting and studying them will increase the knowledge of the scientist. The assumption makes such experiments impossible. I believe that the assumption is sufficient for empirical (or any non-experimental) sciences.

The assumption also eliminates self-selection. For example, one can imagine that students ask about their grades only if they are hopeful of getting a positive grade. However, allowing for a possibility of self-selection should not affect any of the results of this paper. Notice that, if the DM cares only about predictions, self-selection, if biased in a consistent way, only helps: it adds an additional possibility of inference of an outcome from the fact of the query.

Multiple dimensions of features. Suppose that the space of instances is equal to $X = X^1 \times ... \times X^D$, where X_D is infinite. So far I have assumed that $D = 2$. When $D > 2$, an adequate version of the main result of this paper still holds. In particular, the categorization algorithm categorizes not only each of D features, but also pairs of features, triples, ..., and $(D-1)$-tuples of features.

References

Aldous, D. (1981): "Representations for Partially Exchangeable Arrays of Random Variables," *Journal of Multivariate Analysis*, 11, 581–598. 7, 28

Azrieli, Y., and E. Lehrer (2007): "Categorization Generated by Prototypes - An Axiomatic Approach," *Journal of Mathematical Psychology*, 51, 14–28. 4

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler (2005): "Probabilities as similarity-weighted frequencies," *Econometrica*, 73(4), 1125–1136. 5

Blackwell, D., and L. Dubins (1962): "Merging of Opinions with Increasing Information," *Annals of Mathematical Statistics*, 33, 882–886. 15

Canny, J. (2002): "Collaborative Filtering with Privacy Via Factor Analysis," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 6

de Finetti, B. (1964): "La Prevision: Ses Lois Logiques, Ses Sources Subjectives"," in *Studies in Subjective Probability*, ed. by H. E. J. Kybourg, and H. E. Smokler. John Wiley and Sons, New York, translation from French. 7

Gigerenzer, G., and P. M. Todd (1999): *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford. 5

Gilboa, I., O. Lieberman, and D. Schmeidler (2005): "Empirical Similarity," Cowles Foundation Discussion Papers. 5

Gilboa, I., A. Postlewaite, and D. Schmeidler (2004): "Rationality of Belief. Or Why Bayesianism is Neither Necessary Nor Sufficient for Rationality," PIER Working Paper. 3

Gilboa, I., and D. Schmeidler (1989): "Maxmin Expected Utility with Non-Unique Priors," *Journal of Mathematical Economics*, 18, 141–153. 16

Gilboa, I., and D. Schmeidler (1995): "Case-Based Decision Theory," *Quarterly Journal of Economics*, 110(3), 605–639. 2, 5

——— (2000): "Case-based knowledge and induction," *Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans*, 30(2), 85–95. 5

Gilboa, I., and D. Schmeidler (2001): *A Theory of Case-Based Decisions*. University Press, Cambridge, UK. 5

Gilboa, I., and D. Schmeidler (2002): "Cognitive foundations of probability," *Mathematics of Operations Research*, 27(1), 65–81. 5

Glazer, J., and A. Rubinstein (2001): "Debates and Decisions: On a Rationale of Argumentation Rules," *Games and Economic Behavior*, vol. 36(2), 158–173. 19

Hoover, D. (1982): *Row-Column Exchangeability and a Generalized Model for Probability*North-Holland. 7, 28

Hume, D. (1777): *An Enquiry Concerning Human Understanding*. London. 2

Jackson, M. O., and R. G. Fryer Jr (2005): "Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making," Discussion paper. 4, 14

Jackson, M. O., E. Kalai, and R. Smorodinsky (1999): "Bayesian Representation of Stochastic Processes under Learning: De Finetti Revisited," *Econometrica*, 67, 875–894. 15

Jaynes, E. T. (1988): "How Does the Brain Do Plausible Reasoning?," in *Maximum-Entropy and Bayesian Methods in Science and Engineering*, ed. by G. J. Erickson, and C. R. Smith, p. 1. Kluwer, Dordrecht. 2

Jehiel, P. (2005): "Analogy-Based Expectation Equilibrium," *Journal of Economic Theory*, 123, 81–104.

Jehiel, P., and D. Samet (2007): "Valuation Equilibrium," *Theoretical Economics*, 2.

Kallenberg, O. (2005): *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications. Springer, New York. 7, 28

Keynes, J. M. (1921): *A Treatise on Probability*. Macmillan, London. 2

Kleinberg, J., and M. Sandler (2003): "Convergent Algorithms for Collaborative Filtering," *Proceedings of the 4th ACM conference on Electronic commerce*, pp. 1–10. 6

——— (2004): "Using Mixture Models for Collaborative Filtering," *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 569–578. 6

Kreps, D. M. (1988): *Notes on the Theory of Choice*. Westview Press, Boulder. 3

Kumar, R., P. Raghavan, S. Rajagopalan, and A. Tomkins (1998): "Recommendation Systems: A Probabilistic Analysis," *Foundations of Computer Science*, 8-11, 664–673. 6

Lam, R. (2000): "Learning Through Stories," Yale University Ph.D. Thesis. 4

Lehrer, E., and R. Smorodinsky (1996): "Compatible Measures and Merging," *Mathematics of Operations Research*, 21, 697–706. 15, 29

——— (2000): "Relative Entropy in Sequential Decision Problems," *Journal of Mathematical Economics*, 33, 425–439. 15, 29

MARLIN, B., AND R. S. ZEMEL (2004): "The Multiple Multiplicative Factor Model for Collaborative Filtering," in *Proceedings of the Twenty-First International Conference on Machine Learning*. 6

MILL, J. S. (1874): *A System of Logic: Ratiocinative and Inductive*. Harper, New York. 2

MULLAINATHAN, S. (2002): "Thinking Through Categories," NBER and MIT Working Paper. 4

OSHERSON, D. N., O. WILKIE, E. E. SMITH, A. LOPEZ, AND E. SHAFIR (1990): "Category-Based Induction," *Psychological Review*, 97, 185–200. 2, 4, 17, 19, 21

PESKI, M. (2006): "Categorization," University of Chicago, working paper, http://home.uchicago.edu/ mpeski/learning.pdf. 22

RIPS, L. J. (1975): "Inductive Judgements About Natural Categories," *Journal of Verbal learning and Verbal Behavior*, 14, 665–681. 4, 21

SAVAGE, L. J. (1972): *The Foundations of Statistics*. Dover Publications, Toronto. 2

SEGARAN, T. (2007): *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O'Reilly Media, Inc. 5

SMITH, E. E. (1995): "Concepts and Categorization," in *An Invitation to Cognitive Science, Vol.*, ed. by D. N. Osherson, chap. 1, pp. 3–33. MIT Press, Cambridge. 5

TVERSKY, A. (1977): "Features of Similarity," *Psychological Review*, 84, 327–352. 2, 4

TVERSKY, A., AND D. KAHNEMAN (1974): "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185, 1124–1131. 4

APPENDIX A. PROOF OF PROPOSITION 2

The proof relies on a well-known connection between entropy and Bayesian updating. Some preliminary remarks are needed. Observe that

$$\begin{aligned}
&\prod\nolimits_{(x,y)\in d}\rho\left(c\left(x\right)\right)\left(y\right)\\
&=\prod_{k^1,k^2}\left(\left(\rho\left(k^1,k^2\right)\left(1\right)\right)^{\phi\left(k^1,k^2|c,d\right)}\left(\rho\left(k^1,k^2\right)\left(0\right)\right)^{\left(1-\phi\left(k^1,k^2|c,d\right)\right)}\right)^{n\left(k^1,k^2|c,d\right)}\\
&=\exp\left(\sum_{k^1,k^2}n\left(k^1,k^2|c,d\right)\left(\begin{array}{c}\phi\left(k^1,k^2|c,d\right)\log\left(\rho\left(k^1,k^2\right)\left(1\right)\right)\\+\left(1-\phi\left(k^1,k^2|c,d\right)\right)\log\left(\rho\left(k^1,k^2\right)\left(0\right)\right)\end{array}\right)\right)\\
&\leq\exp\left(-\left|d\right|E\left(c|d\right)\right),
\end{aligned}$$

because the expression in the third line is maximized when $\rho\left(k^1,k^2\right)\left(1\right)=\phi\left(k^1,k^2|c,d\right)$. For any database d, and any assignment c, let $C\left(c|d\right)$ be the set of all assignments that coincide

with c on all observations in d :

$$C(c|d) := \left\{c' \in \mathcal{C}^k : c(x) = c'(x) \ \forall (x,y) \in d\right\}.$$

Sets $C(c|d)$ induce partition space $\mathcal{C}^k$ into exactly $k^{d^1+d^2}$ disjoint sets. By the definition of the uniform distribution Ψ^k_C, for any assignments c and c',

$$\Psi^k_C(C(c|d)) = k^{-d^1-d^2}.$$

Also, note that $\psi^k(c,\rho|d)$ depends on assignment c only up to the observations in database d : for any $c' \in C(c|d)$, any $\rho \in \mathcal{R}^k$,

$$\psi^k(c,\rho|d) = \psi^k(c',\rho|d).$$

By the above,

$$\int_{\{c:E(c|d_t)>E_{\min}(d_t)+\varepsilon\}\times\mathcal{R}^k} \prod\nolimits_{(x,y)\in d_t} \rho(c(x))(y)\, d\Psi^k(c,\rho) \leq \exp(-tE_{\min}(d_t) - \varepsilon t).$$

Find an assignment $c_{\max}$ that minimizes the entropy in database d_t, i.e. $E(c_{\max}|d_t) = E_{\min}(d)$. Define prediction function $\rho_{\max}$ such that for each category k^1, k^2,

$$\rho_{\max}\left(k^1,k^2\right)(1) = \phi\left(k^1,k^2|c_{\max},d_t\right).$$

Denote the set of prediction functions

$$\mathcal{R}_t : \left\{\rho \in \mathcal{R}^k : \forall_{k^1,k^2}\forall_y \rho\left(k^1,k^2\right)(y) \geq e^{-\frac{\varepsilon}{2}}\rho_{\max}\left(k^1,k^2\right)(y)\right\}.$$

By the definition of the uniform distribution Ψ^k_R,

$$\Psi^k_R(\mathcal{R}_t) \geq \left(\frac{1}{2}\left(1-e^{-\frac{\varepsilon}{2}}\right)\right)^{k^2}.$$

Using the above calculations, for any prediction function $\rho \in \mathcal{R}_t$,

$$\prod\nolimits_{(x,y)\in d_t} \rho(c_{\max}(x))(y) \geq \exp\left(-t\left(E_{\min}(d_t) + \frac{\varepsilon}{2}\right)\right).$$

Hence,

$$\begin{aligned}
&\int_{\{c:E(c|d_t)\leq E_{\min}(d_t)+\varepsilon\}\times\mathcal{R}^k} \prod\nolimits_{(x,y)\in d_t} \rho(c(x))(y)\, d\Psi^k(c,\rho) \\
&\geq \int_{\{c\in C(c_{\max}|d)\}\times\mathcal{R}_t} \prod\nolimits_{(x,y)\in d_t} \rho(c(x))(y)\, d\Psi^k(c,\rho) \\
&\geq \exp\left(-t\left(E_{\min}(d_t) + \frac{\varepsilon}{2}\right)\right) k^{-d^1_t-d^2_t}\left(\frac{1}{2}\left(1-e^{-\frac{\varepsilon}{2}}\right)\right)^{k^2}.
\end{aligned}$$

Observe that

$$\begin{aligned}
&\lim_{t\to\infty}\frac{\int\limits_{\{c:E(c|d_t)\leq E_{\min}(d_t)+\varepsilon\}\times\mathcal{R}^k}\psi^k\left(c,\rho|d_t\right)d\Psi^k\left(c,\rho\right)}{\int\limits_{\{c:E(c|d_t)>E_{\min}(d_t)+\varepsilon\}\times\mathcal{R}^k}\psi^k\left(c,\rho|d_t\right)d\Psi^k\left(c,\rho\right)}\\
&=\lim_{t\to\infty}\frac{\int\limits_{\{c:E(c|d_t)\leq E_{\min}(d_t)+\varepsilon\}\times\mathcal{R}^k}\prod_{(x,y)\in d_t}\rho\left(c\left(x\right)\right)\left(y\right)d\Psi^k\left(c,\rho\right)}{\int\limits_{\{c:E(c|d_t)>E_{\min}(d_t)+\varepsilon\}\times\mathcal{R}^k}\prod_{(x,y)\in d_t}\rho\left(c\left(x\right)\right)\left(y\right)d\Psi^k\left(c,\rho\right)}\\
&\geq\left(\frac{1}{2}\left(1-e^{-\frac{\varepsilon}{2}}\right)\right)^{k^2}\lim_{t\to\infty}\frac{\exp\left(-t\left(E_{\min}\left(d_t\right)+\frac{\varepsilon}{2}\right)\right)k^{-d_t^1-d_t^2}}{\exp\left(-tE_{\min}\left(d_t\right)-\varepsilon t\right)}\\
&=\left(\frac{1}{2}\left(1-e^{-\frac{\varepsilon}{2}}\right)\right)^{k^2}\lim_{t\to\infty}\exp\left(t\left(\frac{\varepsilon}{2}-\frac{\left(d_t^1+d_t^2\right)}{t}\log k\right)\right)=\infty,
\end{aligned}$$

where the limit is a consequence of the sufficient data condition. This finishes the proof of the Proposition.

Appendix B. Representation of symmetric distributions

B.1. **Representation theorem.** In this part of the Appendix, I present a useful representation of a symmetric distribution. This can be described as follows. For any symmetric ω there exists a measurable function $q:[0,1]^3\to\Delta\{0,1\}$ which can be used to generate the state of the world in the following procedure:

- draw variable $\xi^{\varnothing}$ from the uniform measure on interval $[0,1]$,
- for each feature i, for any $x^i\in X^i$, draw independently $\xi^i\left(x^i\right)$ from measure $U[0,1]$,
- for each instance $(x^1,x^2)\in X$, draw independently $\theta\left(x^1,x^2\right)$ from distribution $q\left(\xi^{\varnothing},\xi^1\left(x^1\right),\xi^2\left(x^2\right)\right.$

The Representation Theorem says that the distribution of θ generated in such a procedure is equal to ω. Variables $\xi^i\left(x^i\right)$ are interpreted as categories of features x^i, or, alternatively, shocks to the outcomes of instances that are feature specific. Variable $\xi^{\varnothing}$ can be interpreted as an aggregate shock to the outcomes of all instances.

Formally, let $\Xi:=[0,1]\times[0,1]^{X^1}\times[0,1]^{X^2}$. I refer to Ξ as the space of auxiliary variables with a typical element $\xi\in\Xi$. For any $x\in X$, denote

$$\xi\left(x\right)=\left(\xi^{\varnothing},\xi^1\left(x^1\right),\xi^2\left(x^2\right)\right)\in[0,1]^3.$$

$\xi\left(x\right)$ is equal to a triple of an auxiliary variable $\xi^{\varnothing}$, an auxiliary variable assigned to feature x^1 and an auxiliary variable assigned to feature x^2. Let λ be a measure on Ξ defined as the product of independent uniform measures on the interval $[0,1]$.

Theorem 2 (Representation of Invariant Distributions)**.** *For any symmetric distribution* $\omega \in \Delta\Theta$, *there is a distribution* $\omega^* \in \Delta(\Theta \times \Xi)$, *such that*

(1) $\text{marg}_\Theta \omega^* = \omega$;
(2) $\text{marg}_\Xi \omega^* = \lambda$;
(3) *there exists a measurable function* $q : [0,1]^3 \to \Delta\{0,1\}$, *such that for any* x

$$\begin{aligned} q(\xi(x)) &= \omega^*(\theta(x) | \xi(x)) \\ &= \omega^*\left(\theta(x) | \xi, \{\theta(x')\}_{x' \neq x}\right), \end{aligned}$$

i.e., conditional distribution of $\theta(x)$, *conditional on the realization of all auxiliary variables* ξ *and all other outcomes* $\theta(x'), x' \neq x$, *depends only on the realization of* $\xi(x)$.

Proof. The Theorem is a restatement of Corollary 7.23 of Kallenberg (2005). This result was originally proven in Aldous (1981) and Hoover (1982). □

Say that measure ω^* *represents* distribution ω. The representing measure is not unique, but the choice of representation is not important for the proof as long as it is fixed. From now on, instead of writing ω^*, I always write ω. The second property says that variables

$$\xi^\varnothing, \xi^1\left(x^1\right)_{x^1 \in X^i}, \xi^2\left(x^2\right)_{x^2 \in X^i}$$

are independent and uniformly distributed on the interval $[0,1]$. The third property says that, conditional on the realization of $\xi(x)$, no additional information (apart from observing outcome $\theta(x)$ itself) affects the prediction of outcome $\theta(x)$. In other words, variable $\xi(x)$ is a *sufficient statistic* for outcome $\theta(x)$.

In order to shorten the notation, write $E_\theta, E_\xi, E_{\theta|\xi}$ to denote expectations with respect to $\theta \in \Theta, \xi \in \Xi$, and, θ conditional on the realization of ξ. In particular,

$$E_{\theta,\xi} = E_\xi E_{\theta|\xi}.$$

B.2. **Approximation.** By the above Theorem, each symmetric distribution can be represented by infinitely many categories from interval $[0,1]$. It turns out that each such a distribution can be approximated by distributions generated only with finitely many categories.

Divide $[0,1)$ into k intervals of equal length and for any $z \in [0,1)$, let $A^k(z) \in \{1,...,k\}$ be the index of the interval that covers z, $z \in \left[\frac{A^k(z)-1}{k}, \frac{A^k(z)}{k}\right]$. For any $(z_0, z_1, z_2) \in [0,1]^3$, define

$$q^k(z_0, z_1, z_2) := E_{z_1', z_2'}\left(q(z_0, z_1', z_2') | A^k(z_i) = A^k(z_i') \text{ for } i = 1,2\right),$$

where the expectation is taken with respect to the uniform measure on $[0,1]^2$. Hence, $q^k(z_0,z_1,z_2)$ is equal to the expectation of $q(z_0,z_1',z_2')$ with respect to i.i.d. uniformly distributed z_1' and z_2', conditional on the fact that $A^k(z_i)=A^k(z_i')$ for $i=1,2$.

Let the expected difference between $q(z)$ and $q^k(z)$ be denoted as

$$\Delta^k := E_z\left\|q^k(z)-q(z)\right\|, \tag{B.1}$$

where the expectation is taken with respect to the uniform measure on $[0,1]^3$. By standard arguments based on the Martingale Convergence Theorem[11],

$$\lim_{k\to\infty}\Delta^k=0. \tag{B.2}$$

Appendix C. Proofs of Theorem 1

The proof of the Theorem is closely related to the much more general argument from Lehrer and Smorodinsky (1996). In order to facilitate the comparison between the two results, I restate the definitions from Lehrer and Smorodinsky (1996). From now on, fix an instance process (x_t) with the sufficient data condition.

Say that the categorization algorithm *almost weakly merges* to (symmetric) measure ω if

$$\lim_{k\to\infty}\lim_{t\to\infty}\frac{1}{t}\sum\nolimits_{s<t}\left\|l_C(d_s,x_s)-\omega(\theta(x_s)=.|\theta(x_1),...\theta(x_{s-1}))\right\|=0,\ \omega\text{-almost surely.} \tag{C.1}$$

(Here, $\omega(\theta(x_s)=.|\theta(x_1),...\theta(x_{s-1}))$ is the conditional distribution of outcome $\theta(x_s)$ given the realization of outcomes $\theta(x_1),...,\theta(x_s)$.) Additionally, for any two (not necessarily symmetric) distributions $\mu,\mu'\in\Delta\Theta$, define the *relative entropy* of μ' with respect to μ at θ as

$$h_\mu^{\mu'}(\theta)=\lim\inf_{t\to\infty}\frac{1}{t}\log\frac{\mu'(\theta(x_1),...,\theta(x_t))}{\mu(\theta(x_1),...,\theta(x_t))}.$$

It follows from standard arguments that $h_\mu^{\mu'}(\theta)\le 0$, μ-almost surely. Lehrer and Smorodinsky (1996) (Theorem 1; see also Proposition 1 of Lehrer and Smorodinsky (2000)) show that if

$$h_\omega^{\omega_C}(\theta)\ge 0,\ \omega\text{-almost surely,} \tag{C.2}$$

then the categorization algorithm almost weakly merges.

I was not able to establish the almost sure bound (C.2). Instead of almost weak merging, Theorem 1 asserts a weaker notion of the convergence in expectations. Nevertheless, the strategy of the proof is analogous: In order to show the convergence of predictions in the

[11]More precisely, the Martingale Convergence Theorem implies that $\lim_{k\to\infty}\Delta^{2^k}=0$. The convergence (B.2) follows from the fact that functions q^k, q are bounded by 1, which implies that for any $k<m$, $\Delta^m\le\Delta^k+\frac{k}{m}$, which in turn implies that for all k, $\limsup_{m\to\infty}\Delta^m\le\Delta^{2^k}$. I am grateful to the referee for pointing the omission in the argument.

expectation, we establish an expectation version of convergence C.2. The latter step is based on the Representation Theorem 2.

Formally, for any two (not necessarily symmetric) distributions $\mu, \mu' \in \Delta\Theta$, define an alternative ("exectation") version of the relative entropy

$$e_{\mu}^{\mu'}(t) = E_{\mu}\frac{1}{t}\log\frac{\mu'(\theta(x_1),...,\theta(x_t))}{\mu(\theta(x_1),...,\theta(x_t))}.$$

Also, define the t-period average distance between predictions

$$d_{\mu}^{\mu'}(t) = \frac{1}{t}E_{\mu}\sum_{s\leq t}\left\|\mu'(\theta(x_s) = .|\theta(x_1),...,\theta(x_{s-1})) - \mu(\theta(x_s) = .|\theta(x_1),...,\theta(x_{s-1}))\right\|^2.$$

(Recall that $\|p-q\|$ is the L^2-distance between measures $p, q \in \Delta\{0,1\}$). The relationship between the two distances is captyured by the followng result:

Lemma 1. $8e_{\mu}^{\mu'}(t) \leq -d_{\mu}^{\mu'}(t)$.

Next, I use Theorem 2 to find a represention of distribution ω : Assume that ω is defined over outcome functions $\theta \in \Theta$ as well as auxiliary variables ξ, that the distribution over auxiliary variables ξ is equal to λ, and that, conditionally on the realization of ξ, outcomes $\theta(x)$ are independenly drawn from distrbution $q(\xi(x)) \in \Delta\{0,1\}$.

Lemma 2. *For each $\varepsilon > 0$, there exists k such that $\liminf_{t\to\infty} E_{\lambda}d_{\omega(.|\xi)}^{\omega^k}(t) \geq -\varepsilon$. Also, $\liminf_{t\to\infty} E_{\lambda}d_{\omega(.|\xi)}^{\omega_C}(t) = 0$. (Here, E_{λ} is the expectation with respect to the uniform measure on the auxiliary variables λ.)*

Lemma 3. *Conditionally on λ-almost all realizations of auxiliary variables ξ, $lim_{t\to\infty}d_{\omega(.|\xi)}^{\omega_C}(t) = 0$.*

It follows from Lemmas 1 and 3that for any measure μ, conditionally on λ-almost all realizations of auxiliary variables ξ

$$\begin{aligned}\limsup_{t\to\infty}\left(d_{\omega}^{\mu}(t)\right)^2 &\leq -8\lim\inf_{t\to\infty}e_{\omega}^{\mu}(t) = -8\lim\inf_{t\to\infty}\left(e_{\omega(.|\xi)}^{\mu}(t) - e_{\omega(.|\xi)}^{\omega}(t)\right)\\ &= -8\lim\inf_{t\to\infty}e_{\omega(.|\xi)}^{\mu}(t).\end{aligned}$$

The Theorem follows from the above inequality and Lemma 2.

C.1. **Proof of Lemma 1.** We start with a preliminary result:

Lemma 4. *For any $p, q \in \Delta\{0,1\}$,*

$$\sum_{y=0,1}q(y)(\log p(y) - \log q(y)) \leq -\frac{1}{8}\|p-q\|^2.$$

Proof. Fix $x \in (0,1)$. Consider function

$$f_x(d) = x\log(x+d) + (1-x)\log(1-x-d) - x\log x - (1-x)\log(1-x)$$

for $-x < d < 1-x$. Then, $f(0) = f'(0) = -0$, and $f'(d) \leq -\frac{1}{2}d$. It follows that $f(d) \leq -\frac{1}{4}d^2$. Now, notice that

$$\begin{aligned}&\sum_{y=0,1} q(y)(\log p(y) - \log q(y))\\ &\leq f_{q(0)}(p(0) - q(0)) \leq -\frac{1}{4}(p(0) - q(0))^2 = -\frac{1}{8}\|p-q\|^2.\end{aligned}$$

□

By the above Lemma, for each s,

$$\begin{aligned}&E_{\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))}\begin{pmatrix}\log\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))\\ -\log\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))\end{pmatrix}\\ &\leq -\frac{1}{8}\|\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1})) - \mu(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))\|^2.\end{aligned}$$

Hence, by the Law of Iterated Expectations and the Jensen's nequality,

$$\begin{aligned}e_\mu^{\mu'}(t) &= \frac{1}{t}E_\mu\log\frac{\mu'(\theta(x_1),...,\theta(x_t))}{\mu'(\theta(x_1),...,\theta(x_t))}\\ &= \frac{1}{t}E_\mu\sum_{s\leq t}E_{\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))}\begin{pmatrix}\log\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))\\ -\log\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))\end{pmatrix}\\ &\leq -\frac{1}{8}E_\mu\sum_{s\leq t}\|\mu'(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1})) - \mu(\theta(x_s)=.|\theta(x_1),...,\theta(x_{s-1}))\|^2\\ &= -\frac{1}{8}d_\mu^{\mu'}(t).\end{aligned}$$

C.2. **Proof of Lemma 2.** We start with the first part of the result. Take any $\varepsilon > 0$ and find $\delta > 0$ so that

$$2\delta\log\delta - 100\delta \geq -\varepsilon.$$

Recall the defintion of the approximation q^k and Δ^k from Appendix B.2. Fnd k sufficiently high so that

$$\sqrt{\Delta^k} < \delta. \tag{C.3}$$

For each realization $\xi^\varnothing \in [0,1]$, define

$$B(\delta;\xi^\varnothing) = \left\{\rho\in\mathcal{R}^k : \begin{array}{c}\sup_{z_1,z_2\in[0,1]}\left\|q^k(\xi^\varnothing,z_1,z_2) - \rho\left(A^k(z_1),A^k(z_2)\right)\right\| \leq 2\delta,\\ \inf_{z_1,z_2\in[0,1]}\inf_y \rho\left(A^k(z_1),A^k(z_2)\right)(y) \geq \delta.\end{array}\right\}.$$

$B(\delta)$ is the set of prediction functions ρ such that prediction of $q^k(\xi^\varnothing,.,.)$ and $\rho\left(A^k(.),A^k(.)\right)$ differ by at most 2δ and that ρ assigns probability at least δ to each outcome. Because $q^k(\xi^\varnothing,.,.)$ is a step function that is constant on the parttion of $[0,1]^2$ into k^2 "squares," and because Ψ_R^k is the uniform measure on $\mathcal{R}^k$, the Ψ_R^k-probability of set $B(\delta;\xi^\varnothing)$ is bounded from below by

$$\Psi_R^k\left(B\left(\delta;\xi^\varnothing\right)\right)\geq\delta^{k^2}. \tag{C.4}$$

For each x, define $c^*(x)=\left(A^k(\xi_1(x)),A^k(\xi_2(x))\right)$. Then, $c^*\in\mathcal{C}^k$. For each database d, let $C(d)$ be the set of assignments that coincide with $c\in\mathcal{C}^k$ on all observations in d (recall the definition from the proof of Proposition 2). The "prior" probability of set $C(d)$ is bounded from below by

$$\Psi_C^k(C(d))\geq\left(\frac{1}{k}\right)^{d^1+d^2}, \tag{C.5}$$

where d^1+d^2 is the total number of distinct features present in the database d. Inequalities (C.4) and (C.5) imply that

$$\Psi^k\left(B\left(\delta;\xi^\varnothing\right)\times C(d)\right)\geq\delta^{k^2}\left(\frac{1}{k}\right)^{d^1+d^2},$$

and that for each t,

$$\begin{aligned}
&\log\omega^k(\theta(x_1),...,\theta(x_t))\\
&\geq\log\left(\Psi^k(B(\delta)\times C(d_t))\right)+\min_{\rho\in B(\delta;\xi^\varnothing),c\in C(d_t)}\sum_{s\leq t}\log\rho(c_1(x_s),c_2(x_s))(\theta(x_s))\\
&\geq\left(d_t^1+d_t^2\right)\log\frac{1}{k}+k^2\log\delta\\
&+\sum_{s\leq t}\min_{\rho\in B(\delta;\xi^\varnothing)}\log\left[\rho\left(A^k(\xi_1(x_s)),A^k(\xi_2(x_s))\right)(\theta(x_s))\right].
\end{aligned} \tag{C.6}$$

Define random variable

$$I_t=\mathbf{1}\left\{\left\|q^k(\xi(x_t))-q(\xi(x_t))\right\|>\delta\right\}.$$

Because of the definition of δ (and Δ^k), the expectation of I_t is bounded by

$$E_\omega I_t\leq\delta.$$

For each s and y, let

$$\rho_s\left(y;\xi^\varnothing\right)=\min_{\rho\in B(\delta;\xi^\varnothing)}\left[\rho\left(A^k(\xi_1(x_s)),A^k(\xi_2(x_s))\right)(y)\right]$$

Then,

$$\rho_s\left(y;\xi^{\varnothing}\right)\geq\delta,\text{ and}$$
$$q\left(\xi\left(x_s\right)\right)\leq\rho_s\left(y;\xi^{\varnothing}\right)+3\delta\text{ whenever }I_t=0.$$

In particular, if $E_{\omega(.|\xi,d_t)}$ is the expectation with respect to distribution ω conditionally on the realization of auxiliary variables ξ and database d_t, then

$$\begin{aligned}
&E_{\omega(.|\xi,d_t)}\left(\log\rho_t\left(\theta\left(x_t\right);\xi^{\varnothing}\right)-\log q\left(\xi\left(x_t\right)\right)\left(\theta\left(x_t\right)\right)\right)\\
&\geq 2I_t\log\delta+\left(1-I_t\right)\sum_{y=0,1}q\left(\xi\left(x_s\right)\right)(y)\left[\log\frac{\rho_t\left(\theta\left(x_t\right);\xi^{\varnothing}\right)}{q\left(\xi\left(x_t\right)\right)\left(\theta\left(x_t\right)\right)}\right]\\
&\geq 2I_t\log\delta-\sum_{y=0,1}\left(\rho_s\left(y;\xi^{\varnothing}\right)+3\delta\right)\log\left(1+\frac{3\delta}{\rho_s\left(y;\xi^{\varnothing}\right)}\right)\\
&\geq 2I_t\log\delta-\sum_{y=0,1}\left(\rho_s\left(y;\xi^{\varnothing}\right)+3\delta\right)\left(\frac{3\delta}{\rho_s\left(y;\xi^{\varnothing}\right)}\right)\\
&\geq 2I_t\log\delta-2\left(3\delta+9\delta\right)\\
&\geq 2I_t\log\delta-100\delta. \qquad \text{(C.7)}
\end{aligned}$$

(Notice that random variable I_t is determined by the realization of ξ and database d_t.)

Because of (C.6), (C.7), the sufficient data condition, and the choice of δ,

$$\begin{aligned}
&\liminf_{t\to\infty}E_\lambda e^{\omega^k}_{\omega(.|\xi)}\\
&=\liminf_{t\to\infty}\frac{1}{t}E_\omega\left(\log\omega^k\left(\theta\left(x_1\right),...,\theta\left(x_t\right)\right)-\sum_{s\leq t}\log q\left(\xi\left(x_s\right)\right)\left(\theta\left(x_s\right)\right)\right) \qquad \text{(C.8)}\\
&\geq\liminf_{t\to\infty}\frac{1}{t}E_\omega\sum_{s\leq t}\left(\log\rho_s\left(\theta\left(x_s\right);\xi^{\varnothing}\right)-\log q\left(\xi\left(x_s\right)\right)\left(\theta\left(x_s\right)\right)\right)\\
&\geq\liminf_{t\to\infty}\frac{1}{t}E_\omega\sum_{s\leq t}\left(2I_t\log\delta-100\delta\right)\geq 2\delta\log\delta-100\delta\geq-\varepsilon. \qquad \text{(C.9)}
\end{aligned}$$

The second part of the Theorem follows from the first part and the fact that because $\omega_C=\alpha_k\omega^k+\left(1-\alpha_k\right)\omega_{-k}$ for some $\alpha_k>0$ and some probability distribution ω_{-k},

$$e^{\omega_C}_{\omega(.|\xi)}\left(t\right)\geq\frac{\log\alpha_k}{t}+e^{\omega^k}_{\omega(.|\xi)}\left(t\right).$$

C.3. **Proof of Lemma 3.** Because $q\left(\xi\left(x_s\right)\right)$ is the conditional distribution of outcome $\theta\left(x_s\right)$ given ξ and $\theta\left(x_1\right),...,\theta\left(x_{s-1}\right)$, and $\omega\left(\theta\left(x_s\right)=.|\theta\left(x_1\right),...,\theta\left(x_{s-1}\right)\right)$ is the conditional

distribution given $\theta(x_1), ..., \theta(x_{s-1})$, standard convexity arguments show that

$$\begin{aligned}&\omega(\theta(x_s) = .|\theta(x_1), ..., \theta(x_{s-1}))\\ &\in \arg\max_{f:(X\times Y)^{s-1}\to\Delta\{0,1\}} E_\omega \left\| q(\xi(x_s))(\theta(x_s)) - f(d_s)(\theta(x_s))\right\|.\end{aligned}$$

The claim follows from Lemma 1.

C.4. **Proof of Theorem 1.** The second part of Theorem follows from the above Lemmas. The proof of the first part of the Theorem follows from almost the same argument. The only difference is that inequality (C.6) is replaced by its analog that does not contain $\log \alpha_k$ and that inequalities (C.9) can be restated with ω^k instead of ω_C and they contain additional term that disappears with k.

Appendix D. Proof of Proposition 3

The proof relies on two Lemmas:

Lemma 5. *There exists $x \geq 0$, such that*

$$\frac{\omega(\theta(o,p) = \theta(o_A,p), PS)}{\omega(PS)} = \frac{\omega(\theta(o,p) = \theta(o_A,p), PD) + x}{\omega(PD) + 2x}.$$

Proof. The proof relies on the Representation Theorem. Let $\omega \in \Delta(\Theta \times \Sigma)$ be a representing measure and let $q : [0,1]^3 \to \Delta\{0,1\}$ be a prediction function from the Theorem. Denote an auxiliary variable

$$V = \left\{\xi^\varnothing, \xi_P(p), \xi_P(p')\right\}.$$

Variable V is distributed uniformly on $[0,1]^3$. For any realization of V, any $y^* \in \{0,1\}$, denote $p_V, q_V^{y^*} \in \Delta\{0,1\}$: for any $y \in \{0,1\}$,

$$\begin{aligned}p_V(y) &:= E_{\xi_O} q\left(\xi^\varnothing, \xi_P(p'), \xi_O\right)(y),\\ q_V(y|y^*) &:= \frac{E_{\xi_O}\left(q\left(\xi^\varnothing, \xi_P(p'), \xi_O\right)(y^*)\right)\left(q\left(\xi^\varnothing, \xi_P(p), \xi_O\right)(y)\right)}{E_{\xi_O(o_A)} q\left(\xi^\varnothing, \xi_P(p'), \xi_O\right)(y^*)},\\ q_V(y) &:= \sum_{y^*} p_V(y^*) q_V(y|y^*),\\ d_V(y|y^*) &:= q_V(y|y^*) - q_V(y).\end{aligned}$$

Here, $p_V(y^*)$ is the probability that the outcome of (o_A, p') is equal to y^* conditional on the realization of variable V; $q_V(y|y^*)$ is the probability that the outcome of (o_A, p) is equal to y conditional on the outcome of (o_A, p') being equal to y^* and the realization of V; $q_V(y)$ is the probability that outcome of (o_A, p) is equal to y conditional on the realization of V. For any function $f : [0,1]^3 \to R$, let

$$E^* f(V) := E_V f(V) p_V^2(0) p_V^2(1),$$

where E_V is an expectation with respect to the realization of variable $V \in \{0,1\}$ (hence, with respect to the uniform measure on $[0,1]^3$).

The subsequent observations are useful in the rest of the proof:

$$\sum_y q_V(y) = 1 \text{ and } d_V(0|y^*) + d_V(1|y^*) = 0 \text{ for each } y^* = 0,1, \tag{D.1}$$

$$\begin{aligned}\sum_{y,y^*} q_V^2(y^*)(y) &= \sum_{y,y^*} (q_V(y) + d_V(y|y^*))^2 \\ &= 2\sum_y q_V^2(y) + 2\sum_{y^*} d_V^2(0|y^*) + 2\sum_{y,y^*} q_V(y) d_V(y|y^*), \end{aligned} \tag{D.2}$$

$$\begin{aligned}\sum_{y,y^*} q_V^2(y|y^*) q_V(y) &= \sum_{y,y^*} (q_V(y) + d_V(y|y^*))^2 q_V(y) \\ &= 2\sum_y q_V^3(y) + \sum_{y^*} d_V^2(0|y^*) \left(\sum_y q_V(y)\right) + 2\sum_{y,y^*} q_V^2(y) d_V(y|y^*) \\ &= 2\sum_y q_V^3(y) + \sum_{y^*} d_V^2(0|y^*) + 2\sum_{y,y^*} q_V^2(y) d_V(y|y^*), \end{aligned} \tag{D.3}$$

$$q_V(y|0) q_V(y|1) = q_V^2(y) + \sum_{y^*} q_V(y) d_V(y|y^*) + d_V(y|0) d_V(y|1). \tag{D.4}$$

Because of (D.1), $d_V(0|0) d_V(0|1) \leq 0$. Denote

$$x := E_V^* \sum_{y^*} \left(d_V^{y^*}(0)\right)^2 - 2E_V^* d_V(0,0) d_V(1,0) \geq 0.$$

By the Representation Theorem and (D.2),

$$\begin{aligned}&\omega(PS) \\ &= \omega(\theta(o_A,p) = \theta(o_B,p), \theta(o_A,p') = \theta(o_B,p') \neq \theta(o_{AC},p') = \theta(o_D,p')) \\ &= \sum_{y,y^*} \omega(\theta(o_A,p) = \theta(o_B,p) = y, y^* = \theta(o_A,p') = \theta(o_B,p') \neq \theta(o_{AC},p') = \theta(o_D,p')) \\ &= E_V p_V^2(0) p_V^2(1) \left(\sum_{y,y^*} q_V^2(y|y^*)\right) \\ &= E_V^* \left(2\sum_y q_V^2(y) + 2\sum_{y^*} d_V^2(0|y^*) + 2\sum_{y,y^*} q_V(y) d_V(y|y^*)\right) \\ &= E_V^* \left(2\sum_y q_V^2(y) + 2\sum_{y,y^*} q_V(y) d_V(y|y^*)\right) + 2E_V^* \sum_{y^*} d_V^2(0|y^*). \end{aligned}$$

By (D.3),

$$\begin{aligned}
&\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right),PS\right)\\
&=\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right)=\theta\left(o_B,p\right),\theta\left(o_A,p'\right)=\theta\left(o_B,p'\right)\neq\theta\left(o_{AC},p'\right)=\theta\left(o_D,p'\right)\right)\\
&=E_Vp_V^2\left(0\right)p_V^2\left(1\right)\left(\sum_{y,y^*}q_V^2\left(y|y^*\right)q_V\left(y\right)\right)\\
&=E_V^*\left(2\sum_y q_V^3\left(y\right)+2\sum_{y,y^*}q_V^2\left(y\right)d_V\left(y|y^*\right)\right)+E_V^*\sum_{y^*}d_V^2\left(0|y^*\right).
\end{aligned}$$

By (D.4),

$$\begin{aligned}
&\omega\left(PD\right)\\
&=\omega\left(\theta\left(o_A,p\right)=\theta\left(o_C,p\right),\theta\left(o_A,p'\right)=\theta\left(o_B,p'\right)\neq\theta\left(o_{AC},p'\right)=\theta\left(o_D,p'\right)\right)\\
&=2\sum_y E_Vp_V^2\left(0\right)p_V^2\left(1\right)q_V\left(y|0\right)q_V\left(y|1\right)\\
&=E_V^*\left(2\sum_y q_V^2\left(y\right)+2\sum_{y,y^*}q_V\left(y\right)d_V\left(y|y^*\right)\right)+4E_V^*d_V\left(0|0\right)d_V\left(0|1\right),
\end{aligned}$$

$$\begin{aligned}
&\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right),PD\right)\\
&=\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right)=\theta\left(o_C,p\right),\theta\left(o_A,p'\right)=\theta\left(o_B,p'\right)\neq\theta\left(o_{AC},p'\right)=\theta\left(o_D,p'\right)\right)\\
&=2E_Vp_V^2\left(0\right)p_V^2\left(1\right)\sum_y q_V\left(y|0\right)q_V\left(y|1\right)q_V\left(y\right)\\
&=E_V^*\left(2\sum_y q_V^3\left(y\right)+2\sum_{y,y^*}q_V^2\left(y\right)d_V\left(y|y^*\right)\right)+2E_V^*d_V\left(0|0\right)d_V\left(0|1\right).
\end{aligned}$$

The above computations lead to the thesis of the Lemma. □

Lemma 6. $2\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right),PD\right)\geq\omega\left(PD\right).$

Proof. By the computations from the previous Lemma,

$$\begin{aligned}
&2\omega\left(\theta\left(o,p\right)=\theta\left(o_A,p\right),PD\right)-\omega\left(PD\right)\\
&=2E_Vp_V^2\left(0\right)p_V^2\left(1\right)\left(\sum_y q_V\left(y|0\right)q_V\left(y|1\right)\left(2q_V\left(y\right)-1\right)\right).
\end{aligned}$$

Because $q_V(0|y^*) = 1 - q_V(1|y^*)$ and $q_V(0) = 1 - q_V(1)$, it must be that

$$\begin{aligned}
&\sum_y q_V(y|0)\, q_V(y|1)\,(2q_V(y) - 1) \\
&= (2q_V(0) - 1)\,(q_V(0|0)\, q_V(0|1) - (1 - q_V(0|0))\,(1 - q_V(0|1))) \\
&= (2q_V(0) - 1)\,(q_V(0|0) + q_V(0|1) - 1).
\end{aligned}$$

□

By the first Lemma, either

$$\frac{\omega(\theta(o,p) = \theta(o_A,p), PS)}{\omega(PS)} \geq \frac{1}{2} \text{ and } \frac{\omega(\theta(o,p) = \theta(o_A,p), PD)}{\omega(PD)} \geq \frac{1}{2} \text{ or}$$
$$\frac{\omega(\theta(o,p) = \theta(o_A,p), PS)}{\omega(PS)} < \frac{1}{2} \text{ and } \frac{\omega(\theta(o,p) = \theta(o_A,p), PD)}{\omega(PD)} < \frac{1}{2}.$$

Because of the second Lemma and the fact that

$$\omega(\theta(o,p) = \theta(o_A,p)) = \omega(\theta(o,p) = \theta(o_A,p), PS) + \omega(\theta(o,p) = \theta(o_A,p), PD),$$

the second pair of inequalities is impossible. But then, by the first Lemma,

$$\begin{aligned}
\omega(\theta(o,p) = \theta(o_A,p)\,|PS) &= \frac{\omega(\theta(o,p) = \theta(o_A,p), PS)}{\omega(PS)} \\
&\leq \frac{\omega(\theta(o,p) = \theta(o_A,p), PD)}{\omega(PD)} = \omega(\theta(o,p) = \theta(o_A,p)\,|PD).
\end{aligned}$$

Department of Economics, University of Texas at Austin, 1 University Station #C3100, Austin, Texas 78712

E-mail address: mpeski@gmail.com