PRIOR SYMMETRY, CATEGORIZATION, AND SIMILARITY-BASED REASONING*

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, then 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 the model.

" ... we conclude [that] because a resembles b in one or more properties, that it does so in a certain other property." (Mill 1874)

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, Jorg Stoye, and Nabil al-Najar 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 suggested a simplification in the proof of Theorem 1. The author bears full responsibility for the remaining errors.

**The University of Chicago, Department of Economics. E-mail: mpeski@uchicago.edu.

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)).

Categorization is a model of similarity-based reasoning (see also (Tversky 1977), (Gilboa and Schmeidler 1995) or (Osherson, Wilkie, Smith, Lopez, and Shafir 1990)). It has three defining characteristics. First, the DM divides objects and properties into a finite number of groups called *categories* considering two objects to be similar if they are assigned to the same category. Thus, similarity has a particularly simple binary form. Second, categories are few (relative to the number of objects) and large. Third, 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 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 a possibility of categorization before the first period because before the first period the DM cannot distinguish which objects are more similar to one another than to the rest.

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) chapters 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 5 constructs a learning algorithm that has the features of the psychological model of categorization: The DM assigns objects into a few large categories and makes category-based predictions. The assignments into categories minimizes the inner-category entropy and it 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 5.3, 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 is large. 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 5.4, 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**. In the long-run, any optimal behavior in symmetric environments is behaviorally indistinguishable from categorization.

This implies that a behavior of the rational DM should look *as if* she categorized, even if she is really following some other optimal learning rule.

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 6, 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 somewhat 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 a tool for making predictions. The evolutionary pressure should have optimized this tool for standard environments. The above results imply that if the standard 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)). (Jehiel 2005) and

(Jehiel and Samet 2007) analyze the behavior of categorizing agents in games. (Azrieli and Lehrer 2004) develops an axiomatic characterization of categorization. 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.

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 because 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 satisfy the combination axiom.[1] Thus, the combination axiom is not necessary for similarity-based reasoning.

3. Model

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 infinite. 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 $x = (x^1, x^2) =$ (student, course). An outcome $\theta(x^1, x^2)$ is equal to 1 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.*

[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.

Example 3 (Recommendation algorithms)**.** *Netflix.com is a DVD rental company. Among numerous services, Netflix helps consumers choose movies using a recommendation algorithm: a customer rates movies, and Netflix uses the ratings of that customer and other customers to predict a rating for the movies that the customer has not yet watched. 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*?"*

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* $l : \bigcup_t (X \times Y)^{t-1} \times X \to \Delta\{0,1\}$ 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

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

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

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

exists, and it is unique and continuous in p.[2] Denote the solution as $a_{\max}(p)$. Let $a : \bigcup_t (X \times Y)^{t-1} \times X \to R$ designate a *behavioral rule*. Let

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

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

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

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.

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

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

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

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, write $l(d, x)$ for database d and an instance x. For example, Bayesian learning rule l_ω does not depend on the order of observations.

Definition 1. *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. Prior symmetry

This section discusses the key assumption of this paper.

4.1. **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 2. *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) and (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.

4.2. **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, 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

[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.

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* $\theta(c,m) = 0$*.*

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 indepentenly 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.

It is also good to give an example of distribution that does not satisfy symmetry. In general, any logical relationship eliminates symmetry.

Example 7. *Consider Example 1 and suppose that property* p *is known to be a logical negation of property* p'*, i.e., if* $\theta(o,p) = 1$ *then* $\theta(o,p') = 0$ *for any object* o*. Then, no distribution that respects this relation is symmetric.*

The results of the paper remain true under assumptions that are somewhat weaker than symmetry (and which apply to Example 7). This is discussed in the last section.

4.3. **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*

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

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,

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

This yields the Proposition. □

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

5. Categorization

5.1. **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: (i.e., 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 $\rho(k^1, k^2) \in [0,1]$. 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 a 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 the space of prediction functions as

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

A couple 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 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 infinitely dimensional space of theories, and it is treated as the prior beliefs.

The posterior density of theory (c,ρ) is defined through Bayes' formula

$$\psi^k(c,\rho|d) = \frac{\prod_{(x,y)\in d}\rho(c(x))(y)}{\int_{\mathcal{T}^k}\prod_{(x,y)\in d}\rho'(c'(x))(y)\,d\Psi^k(c',\rho')},$$

where

$$\prod_{(x,y)\in d}\rho(c(x))(y)$$

is equal to the probability of realization of database d given theory (c,ρ).

Finally, define learning rule l^k as the prediction that would be made by a DM with "beliefs" $\psi(.|d)$:

$$l^k(d,x)(y) = \int_{\mathcal{T}^k}\left[\rho(c(x))(y)\right]\psi^k(c,\rho|d)\,d\Psi^k(c,\rho).$$

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

So far, I have kept the number of categories constant. The second learning rule increases the number of categories together with the size of the data sample. Let K_t de defined as

$$K_t := \left\lfloor \sqrt[3]{\frac{t}{d_t^1+d_t^2}} \right\rfloor, \tag{5.1}$$

where $\lfloor x\rfloor$ is the largest natural number not larger than x. K_t will determine the number of categories used in period t. I refer to K_t as a *switching rule*. If the sufficient data condition (Definition 1) is satisfied, then $\lim_{t\to\infty}K_t=\infty$.

Let $d_t(k)$ be defined as follows:

$$d_t(k) = \{(x_s,y_s) : s<t, K_s=k\}.$$

This is a database of observations before period t in which the value of the switching rule is equal to k. Define an *(adaptive) categorization algorithm* l^C as

$$l^C(d_t,x) := l^{K_t}(d_t(K_t),x).$$

In period t, the DM applies the K_t-categorization algorithm to database $d_t(K_t)$.

Notice that both learning rules satisfy two defining properties of categorization that have been described in the beginning of this Section. The assignment into categories depends on past data. The size of categories grows to infinity as the number of observations increases. (This is because

$$\lim_{t\to\infty}\frac{t}{k}=\infty \text{ and } \lim_{t\to\infty}\frac{t}{K_t}\geq\lim_{t\to\infty}t^{2/3}=\infty.$$

Thus, the categorizing behavior does not disappear in the limit.) Finally, the predictions of outcomes are based on category assignments.

5.2. **Categorization as entropy minimization.** It is instructive 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

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

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 on 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. I will use the Proposition to compare the categorization algorithm from this paper with a heuristics proposed in (Jackson and Fryer Jr 2005). There, the decision maker looks for a categorization assignment that minimizes variance inside categories. Since entropy minimization is not the same as minimization of inner-category variance, (Jackson and Fryer Jr 2005)'s algorithm is not going to satisfy the optimality results that are described next.

5.3. **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. For any $p, q \in \Delta\{0,1\}$, let $\|p-q\| = \sum_{y\in\{0,1\}} |p(y) - q(y)|$ be the L^1-distance between p and q. Let

$$E \left\| l^k(d_t, x_t) - l_\omega(d_t, x_t) \right\|$$

be the expected difference between predictions made in period t by the Bayesian DM and by the k-categorization algorithm, where the expectation is taken with respect to the distribution over past databases induced by ω and instance process $\bar{x}$.

Proposition 3. *Suppose that process $\bar{x}$ satisfies the sufficient data condition. For any symmetric ω,*

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

It is a consequence of the Proposition that 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. The key step in the proof uses the Representation Theorem from Appendix B.1 to show that any symmetric distribution can be approximated as a mixture of distributions generated as in Examples 4, 5 and 6 but with k categories for each feature. Given sufficient data, the categorization algorithm l^k behaves as the Bayesian updating on the approximating distribution. Details can be found in Appendix C.

The Proposition is not satisfactory because the number of categories k needed depends on distribution ω. That, in particular, is not known by the non-Bayesian DM. This is a problem because the DM should choose a learning rule before making any observations and before learning anything about distribution ω.

The DM faces a fundamental trade-off. On one hand, a higher k allows for a better limit approximation, as described in Proposition 3. On the other hand, too high a k creates an *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. The solution to this problem must balance the growth of k with overfitting risks. I use the adaptive categorization

algorithm. In this algorithm, the DM changes the number of categories as time goes by, and the database of past observations increases.[4]

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

$$\limsup_{t\to\infty} \frac{1}{t} E \sum\nolimits_{s<t} \left\| l^C\left(d_s, x_s\right) - l_\omega\left(d_s, x_s\right) \right\| = 0. \tag{5.2}$$

The Theorem says that the adaptive categorization l^C makes, on average, the same predictions as the Bayesian DM, *uniformly* across all symmetric beliefs ω.[5] The Theorem fixes the flaw of Proposition 3: Adaptive categorization does not depend on the initial choice of the number of categories and increases the number of categories together with the data sample. The proof can be found in Appendix C.

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 were 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 high 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, only that the DN will predict as well 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, one can show that if the sufficient data condition is satisfied, then, asymptotically, the categorization algorithm predicts correctly almost all outcomes in Examples 5 and 6.

5.4. **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 an optimal rule is unique. In fact, there are infinitely many optimal learning rules. This is a consequence of my

[4]In general, this method of solving the overfitting problem is known in the statistical decision theory as the *capacity control method* (see, among others, (Vapnik 1998) and (Bousquet, Boucheron, and Lugosi 2004)).

[5]The statement of the Theorem is related to the literature on the Bayesian merging of opinions ((Blackwell and Dubins 1962), (Lehrer and Smorodinsky 2000); also (Jackson, Kalai, and Smorodinsky 1999)). Here and there, the predictions of two learning rules converge only on average. Note, however, that the categorization algorithm l^C is not, strictly speaking, Bayesian, as it is not based on updating one "uniform prior" over all symmetric dstributions. (In fact, such a prior does not exist.)

optimality criterion: Any learning rule that behaves differently from l^C in the first t periods and then follows the same predictions as l^C satisfies the formula (5.2).

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.

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

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

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,

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

with strict inequality if

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

Together with Theorem 1, this implies the thesis of the Corollary. □

6. Similarity-based reasoning

In this section, I discuss a series of laboratory observations and stylized facts about the similarity-based reasoning. In each case, the empirical observation is compared with a Bayesian prediction of the DM with symmetric beliefs. In all cases but the last one, the

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

theoretical predictions coincide with the laboratory observations (the empirical evidence in the last case seems to be quite weak, though).

6.1. **Premise-conclusion similarity.** (Osherson, Wilkie, Smith, Lopez, and Shafir 1990) present subjects with pairs of inductive arguments. Each argument consists of a premise, which is 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. The first argument was chosen as more credible by 59 out of 80 subjects. The interpretation is that a category of sparrows is more similar to categories of robins and bluejays and the inductive argument seems more appropriate.

Note that the careful choice of objects and properties for the experiments ensures that the subjects have "no reason to believe" that the objects and properties are a priori different. (It is rare for non-professionals to be able to differentiate prior probabilities of "using" and "not using serotonin as a neurotransmitter.") Hence, if the Principle of Insufficient Reason holds, then the subjects should have symmetric beliefs. This allows me to treat the coincidence between empirics and theory as an indirect positive verification of the model, or precisely, the fact that the DM is rational and that she has symmetric beliefs. Under these two assumptions, the Principle of Similarity is an implication of Proposition 1.

Next, (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 3. Here, 75 out of 80 subjects point to the second argument as more credible.

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

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

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

TABLE 3. Premise diversity

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 4. Premise diversity

6.2. **Premise diversity.** The next experiment indicates that an inductive argument is more credible if the premise is more diverse. Consider an example in Table 4. Here, 76 out of 80 subjects point to the first argument as more credible.[6]

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 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.

I formalize this argument. Suppose that the set of objects is divided into infinitely many categories, $O = C \times I$, where C is an infinite set of categories and I is an infinite set of identifiers of individual objects. The division into categories can be endogenous, as a result of the past learning, or it might come from some other source of information (such

[6]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.

as instructions given in the experiment). Let $\omega \in \{0,1\}^O$ be the beliefs of the DM about whether objects $o \in O$ have certain fixed property.

I will make two symmetry assumptions about ω: that objects inside each category are exchangeable and that categories are exchangeable. Formally, for each c, for all permutations $\pi : I \to I$, $\pi_C : C \to C$, define permutations $\pi_c^*, \pi_C^* : O \to O$, as

$$\pi_c^*(c', i) = \left\{ \begin{array}{ll} (c', i), & \text{if } c' \neq c, \\ (c', \pi(i)) & \text{if } c' = c, \end{array} \right\},$$
$$\pi_C^*(c, i) = (\pi_C(c), i).$$

Thus, permutation π_C^* exchanges categories and permutation π_c^* exchanges objects inside a particular category c. I assume that for all permutations $\pi : I \to I$, $\pi_C : C \to C$, any category $c \in C$, any measurable set $E \subseteq \{0,1\}^O$,

$$\omega(\theta \in E) = \omega(\pi_c^* \theta \in E) \text{ and } \omega(\theta \in E) = \omega(\pi_C^* \theta \in E).$$

These two assumptions seems appropriate for the experiment in Table 4. Suppose that c_0 is a category of large African mammals, that includes rhinoceroses, hippopotamuses, and, possibly elephants. The first assumption implies that the probability that elephants have the property given that rhinoceroses have it is the same as given hippopotamuses have the property. Unless a DM has biology training, there is no reason why the former should be more informative about the functioning of elephant liver than the latter, and vice versa. Additionally, suppose that c_1 and c_2 are categories of rodents and dogs, respectively. The second assumption says that the probability that dogs have the property is the same, conditionally on either the fact about rodents or large African mammals. Again, there is no reason to expect rodents bring more information about dogs than hippopotamuses.

Proposition 4. *Suppose that ω satisfies the above assumptions. Then, for any objects $(c_0, i_0), (c_1, i_1), (c_2, i_2)$, and (c_0, i_0') such that $c_k \neq c_l$ for $k \neq l$ and $i_0 \neq i_0'$,*

$$\omega(\theta(c_2, i_2) = \theta(c_0, i_0) | \theta(c_0, i_0) = \theta(c_0, i_0'))$$
$$\leq \omega(\theta(c_2, i_2) = \theta(c_0, i_0) | \theta(c_0, i_0) = \theta(c_1, i_1)).$$

The conditioning event in the LHS of the above inequality contains information about two objects from the same category. Thus, it can be interpreted as a *similar premise*. The conditioning event in the RHS contains information about two objects from different categories, and it is interpreted as a *diverse premise*. The Proposition says that the probability that new object has given property increases when the premise is diverse. The proof of the Proposition can be found in Appendix D.

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 5. Premise-conclusion asymmetry

6.3. **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 5. (Osherson, Wilkie, Smith, Lopez, and Shafir 1990) report that a majority of students select the first argument as more credible.[7] Following (Rips 1975), they argue that "mice" are more typical animals; hence, it is more informative about properties shared by a 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: $\mu \in \Delta\left(\{0,1\}^{\{M,B\}}\right)$. Then the conditional probabilities conditional on the premise in the first and second arguments 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.

7. CONCLUDING REMARKS

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

[7]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.

probably the second role of the categorization that is more connected to bounded rationality and biases. A complete theory should take into account both functions of categorization.

There are various possible extensions of the model of learning through categories. 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 5 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 achieve 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.

Finite symmetry. One can also to some degree relax the symmetry assumption. 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

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

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, Theorem 1 holds for any finitely symmetric ω. Thus, adaptive categorization produces predictions asymptotically close to Bayesian even if the distribution over states of the world is finitely symmetric. Corollary 1 follows.

In particular, consider Example 7 and suppose that $P = P^0 \cup P^1$ is an union of two disjoint sets P^0 and P^1. There is a bijection $i : P^0 \to P^1$ such that $i(p)$ is interpreted as the property that is a logical negation of property p. Then, any distribution that is symmetric on $P^0 \times O$ but otherwise respects logical relationships satisfies the above assumptions.

References

Aldous, D. (1981): "Representations for Partially Exchangeable Arrays of Random Variables," *Journal of Multivariate Analysis*, 11, 581–598. 8, 26

Azrieli, Y., and E. Lehrer (2004): "Categorization Generated by Prototypes - An Axiomatic Approach," . 5

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

Bousquet, O., S. Boucheron, and G. Lugosi (2004): "Introduction to Statistical Learning Theory," in *Advanced Lectures in Machine Learning*, ed. by O. Bousquet, U. Luxburg, and G. Rätsch, pp. 169–207. Springer. 15

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. 8

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. 18

HEWITT, E., AND J. L. SAVAGE (1955): "Symmetric Measures on Cartesian Products," *Tran. Amer. Math. Soc.*, 80, 470–501. 34

HOOVER, D. (1982): *Row-Column Exchangeability and a Generalized Model for Probability*North-Holland. 8, 26

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, 13

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. 4

JEHIEL, P., AND D. SAMET (2007): "Valuation Equilibrium," *Theoretical Economics*, 2. 5

KALLENBERG, O. (2005): *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications. Springer, New York. 8, 26

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

KINGMAN, J. F. C. (1978): "Uses of Exchangeability," *Ann. Probability*, 6, 183–197. 34

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

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

LEHRER, E., AND R. SMORODINSKY (2000): "Relative Entropy in Sequential Decision Problems," *Journal of Mathematical Economics*, 33, 425–439. 15

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, 20

PESKI, M. (2006): "Categorization," University of Chicago, working paper, http://home.uchicago.edu/ mpeski/learning.pdf. 21

RIPS, L. J. (1975): "Inductive Judgements About Natural Categories," *Journal of Verbal learning and Verbal Behavior*, 14, 665–681. 4, 20

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

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. 20

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

VAPNIK, V. N. (1998): *Statistical Learning Theory*. Wiley-Interscience. 15

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 :

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

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

$$\Psi_C^k\left(C\left(c|d\right)\right)=k^{-d^1-d^2}.$$

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

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

By the above,

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

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

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

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_R^k,

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

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

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

Hence,

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

Observe that

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

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 $\xi^i(x^i)$ from measure $U[0,1]$,
- for each instance $(x^1, x^2) \in X$, draw independently $\theta(x^1, x^2)$ from distribution $q(\xi^{\varnothing}, \xi^1(x^1), \xi^2(x^2))$

The Representation Theorem says that the distribution of θ generated in such a procedure is equal to ω. Variables $\xi^i(x^i)$ 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

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

$\xi(x)$ 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.

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

1. $\omega = \operatorname{marg}_{\Theta} \omega^*$;
2. $\operatorname{marg}_{\Xi} \omega^*$ *is a product of independent uniform measures on the interval* $[0,1]$;
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 $[0,1]$. One shows that each such distribution can be approximated by distributions generated only with finitely many categories. This part of the Appendix develops the appropriate notation. 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 covering interval: $z \in \left[\frac{A^k(z)-1}{k}, \frac{A^k(z)}{k}\right]$. 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,

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

Given a realization of auxiliary variables ξ, define *approximate* k-assignment, defined as $c^k \in \mathcal{C}^k$, where

$$c^*(x) := \left(A^k\left(\xi^1\left(x^1\right)\right), A^k\left(\xi^2\left(x^2\right)\right)\right).$$

Define *approximate* prediction function $q^{k*} \in \mathcal{R}^k$: for any $k_i \in \{1,...,k\}$

$$\rho^*(k_1, k_2) := q^k\left(\xi^{\varnothing}, \frac{k_1}{k}, \frac{k_2}{k}\right).$$

Both approximate assignment c^* and prediction function ρ^* are defined as functions of ξ and, thus, they are random variables.

Appendix C. Proofs of optimality results

In this part of the Appendix, I present proofs of the optimality results in Section 5.3. The goal is to show that the predictions of categorization algorithms l^k and l_C are close to the prediction of l_ω. This is shown in three steps. First, I show that the period t predictions of the categorization algorithm are close to the prediction of $q\left(\xi\left(x_t\right)\right)$. Next, I show that l_ω is close to $q\left(\xi\left(x_t\right)\right)$. The proofs of Proposition 3 and Theorem 1 follow.

C.1. **Convergence $l^k \to q$.** For any $p, p' \in \Delta\{0,1\}$, let

$$D\left(p, p'\right) = \sum_y p\left(y\right) \log \frac{p\left(y\right)}{p'\left(y\right)}.$$

$D\left(p, q\right)$ measures the distance between p and p'. In particular, $D\left(p, p'\right) = 0$ if and only if $p = p'$.

Proposition 5. *For any instance process $\bar{x}$, for any symmetric ω,*

$$\begin{aligned} &E_{\theta,\xi} \sum_{s<t} D\left(q\left(\xi\left(x_s\right)\right), l^k\left(d_s, x_s\right)\right) \\ &\leq 4\left(\sqrt{\Delta^k} \log \frac{1}{\Delta^k} + 6\sqrt{\Delta^k}\right) t + k^2 \log k + k\left(d_t^1 + d_t^2\right). \end{aligned}$$

(Recall that constant Δ^k is defined in equation (B.2).) The Proposition derives a bound on the distance between prediction $q\left(\xi\left(x_s\right)\right)$ from the Representation Theorem and the prediction of the k-categorization algorithm. The rest of this part of the Appendix proves the Proposition.

C.1.1. *Notation.* For any database d and any assignment c, let $C\left(c|d\right)$ 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). For any $\delta < \frac{1}{2}$, any $\rho \in \mathcal{R}^k$, let

$$B\left(\rho, \delta\right) = \left\{ \rho' \in \mathcal{R}^k : \begin{array}{c} \sup_{\mathbf{k} \in \{1,...,k\}^2} \left\|\rho\left(\mathbf{k}\right) - \rho'\left(\mathbf{k}\right)\right\| \leq 2\delta, \\ \inf_{\mathbf{k} \in \{1,...,k\}^2} \inf_y \rho'\left(\mathbf{k}\right)\left(y\right) \geq \delta. \end{array} \right\}.$$

$B\left(\rho, \delta\right)$ is the set of prediction functions ρ' such that for each category $\mathbf{k} \in \{1, ..., k\}^2$,

- the predictions of ρ and ρ' differ by at most 2δ, and
- ρ' assigns probability at least δ to each outcome.

The "prior" probability of set $B\left(\rho, \delta\right)$ is bounded from below by

$$\Psi_R^k\left(B\left(\rho, 2\delta\right)\right) \geq \delta^{k^2}.$$

Fix a realization of auxiliary variables ξ. Recall that $\psi^k\left(.|d\right)$ denotes the posterior density of the DM's beliefs from the k-categorization algorithm. Let $\Psi\left(.|d\right)$ denote the corresponding

distribution. Also, recall that d_s is a database of observations made before period s. Further denote

$$G(\xi,\delta) := \{(c,\rho) : c \in C(c^*|d_t), \rho \in B(\rho^*,\delta)\},$$
$$l_s^\delta(y) = \frac{1}{\Psi(G(\xi,\delta)|d_s)} \int\limits_{G(\xi,\delta)} \rho(c(x_s))(y)\, d\Psi(c,\rho|d_s),$$

Set $G(\xi,\delta)$ consists of theories (c,ρ) such that

- assignment c agrees with approximate assignment c^* on database d_t, and
- prediction function ρ is close to prediction ρ^*.

Prediction l_s^δ is equal to the prediction of the k-categorization algorithm, conditional on the fact that the theory belongs to set $G(\xi,\delta)$. In particular,

$$\left\|q^k(\xi(x_s)) - l_s^\delta\right\| = \left\|\rho^*(c^*(x_s)) - l_s^\delta\right\| \leq 3\delta.$$

C.1.2. *Intermediary lemmas.*

Lemma 1.

$$E_{\theta,\xi} \sum_{s<t} D\left(q(\xi(x_s)), l_s^\delta\right) \leq 2\left(\frac{\Delta^k}{\delta}\log\frac{1}{\delta} + 6\delta\right)t$$

Proof. Notice that

$$\begin{aligned}
&E_{\theta,\xi} \sum_{s<t} D\left(q(\xi(x_s)), l_s^\delta\right) \\
&\leq E_{\theta,\xi} \sum_{s<t} \mathbf{1}\left\{\left\|q(\xi(x_s)) - q^k(\xi(x_s))\right\| \geq \delta\right\} \max_{p\in\Delta\{0,1\}} D\left(p, l_s^\delta\right) \\
&+ E_{\theta,\xi} \sum_{s<t} \mathbf{1}\left\{\left\|q(\xi(x_s)) - q^k(\xi(x_s))\right\| \leq \delta\right\} D\left(q(\xi(x_s)), l_s^\delta\right).
\end{aligned}$$

To bound the first term, recall that by construction $l_s^\delta(y) \geq \delta$; hence, $D\left(q, l_s^\delta\right) \leq 2\log\frac{1}{\delta}$ for any $q \in \Delta\{0,1\}$. By Chebyshev's inequality,

$$E_{\theta,\xi} \sum_{s<t} \mathbf{1}\left\{\left\|q(\xi(x_s)) - q^k(\xi(x_s))\right\| \geq \delta\right\} \leq t\frac{\Delta^k}{\delta},$$

where Δ^k is defined in equation (B.1).

For the second term, notice that if $\left\|q\left(\xi\left(x_s\right)\right)-q^k\left(\xi\left(x_s\right)\right)\right\| \leq \delta$, then $\left\|q\left(\xi\left(x_s\right)\right)-l_s^\delta\right\| \leq 3\delta$, and

$$\begin{aligned} D\left(q\left(\xi\left(x_s\right)\right), l_s^\delta\right) &\leq \sum_y \left(l_s^\delta\left(y\right)+3\delta\right)\log\left(1+\frac{3\delta}{l_s^\delta\left(y\right)}\right) \\ &\leq 3\delta\sum_y \frac{l_s^\delta\left(y\right)+3\delta}{l_s^\delta\left(y\right)} \leq 12\delta. \end{aligned}$$

The last inequality uses the fact that $l_s^\delta\left(y\right)\geq\delta$ by construction. □

Lemma 2.

$$\begin{aligned} &E_{\theta,\xi}\sum_{s<t}\sum_y q\left(\xi\left(x_s\right)\right)\left(y\right)\log\frac{l_s^\delta\left(y\right)}{l^k\left(d_s,x_s\right)\left(y\right)} \\ &\leq k^2\log\frac{1}{\delta}+k\left(d_t^1+d_t^2\right). \end{aligned}$$

Proof. Notice that for each y,

$$\begin{aligned} &\frac{l_s^\delta\left(y\right)}{l^k\left(d_s,x_s\right)\left(y\right)} \\ &= \frac{1}{\Psi\left(G\left(\xi,\delta\right)|d_s\right)}\frac{\int\limits_{G(\xi,\delta)}\rho\left(c\left(x\right)\right)\left(y\right)d\Psi\left(c,\rho|d_s\right)}{\int\limits_{\mathcal{T}^k}\rho\left(c\left(x\right)\right)\left(y\right)d\Psi\left(c,\rho|d_s\right)} \\ &= \frac{1}{\Psi\left(G\left(\xi,\delta\right)|d_s\right)}\frac{\Psi\left(\theta\left(x_s\right)=y,G\left(\xi,\delta\right)|d_s\right)}{\Psi\left(\theta\left(x_s\right)=y|d_s\right)} \\ &= \frac{\Psi\left(G\left(\xi,\delta\right)|d_s,\theta\left(x_s\right)=y\right)}{\Psi\left(G\left(\xi,\delta\right)|d_s\right)}. \end{aligned}$$

Here, $\left\{d_s,\theta\left(x_s\right)=y\right\}=d_{s+1}$ is a database of past observations in period $t+1$. Hence,

$$\begin{aligned} &\sum_y q\left(\xi\left(x_s\right)\right)\left(y\right)\log\frac{l_s^\delta\left(y\right)}{l^k\left(d_s,x_s\right)\left(y\right)} \\ &= \sum_y q\left(\xi\left(x_s\right)\right)\left(y\right)\log\frac{\Psi\left(G\left(\xi,\delta\right)|d_{s+1}\right)}{\Psi\left(G\left(\xi,\delta\right)|d_s\right)} \\ &= E_{d_{s+1}|d_s,\xi}\log\frac{\Psi\left(G\left(\xi,\delta\right)|d_{s+1}\right)}{\Psi\left(G\left(\xi,\delta\right)|d_s\right)}, \end{aligned}$$

where $E_{d_{s+1}|d_s,\xi}$ is the expectation of the realization of outcome $\theta\left(x_s\right)$, conditional on the realization of outcomes $\theta\left(x_{s'}\right), s'<s$, and auxiliary variables ξ. (The last equality follows from the conditional independence of point 3 of the Representation Theorem.) By the Law

of Iterated Expectations,

$$\begin{aligned}
&E_{\theta,\xi}\sum_{s<t}\sum_{y} q\left(\xi\left(x_s\right)\right)\left(y\right)\log\frac{l_s^{\delta}\left(y\right)}{l^k\left(d_s,x_s\right)\left(y\right)}\\
&=E_{\xi}\left[E_{d_1|\xi}\left[\log\frac{\Psi\left(G\left(\xi,\delta\right)|d_1\right)}{\Psi\left(G\left(\xi,\delta\right)|d_0\right)}+...\left[..+E_{d_t|d_{t-1},\xi}\left[\log\frac{\Psi\left(G\left(\xi,\delta\right)|d_t\right)}{\Psi\left(G\left(\xi,\delta\right)|d_{t-1}\right)}\right]\right]\right]\right]\\
&=E_{d_t,\xi}\log\frac{\Psi\left(G\left(\xi,\delta\right)|d_t\right)}{\Psi\left(G\left(\xi,\delta\right)\right)}\leq -E\log\Psi\left(G\left(\xi,\delta\right)\right)\\
&=-E\log\Psi_C\left(C\left(c^*|d_t\right)\right)-E\log\Psi_R\left(B\left(\rho^*,2\delta\right)\right).
\end{aligned}$$

The thesis of the Lemma is a consequence of the fact, that for any assignment $c\in\mathcal{C}^k$ and for any prediction function $\rho\in\mathcal{R}^k$,

$$\begin{aligned}
\log\Psi_C\left(C\left(c|d\right)\right)&\geq -k\left(d_t^1+d_t^2\right)\text{ and}\\
\log\Psi_R\left(B\left(\rho,2\delta\right)\right)&\geq k^2\log\delta.
\end{aligned}$$

□

C.1.3. *Proof of the Proposition.* Notice that for any $\rho'\in\Delta\left\{0,1\right\}$,

$$D\left(q,p\right)=D\left(q,p'\right)+\sum_{y}q\left(y\right)\log\frac{p'\left(y\right)}{p\left(y\right)}. \tag{C.1}$$

By (C.1) and the Lemmas,

$$\begin{aligned}
&E_{\theta,\xi}\sum_{s<t}D\left(q\left(\xi\left(x_s\right)\right),l^k\left(d_s,x_s\right)\right)\\
&=E_{\theta,\xi}\sum_{s<t}D\left(q\left(\xi\left(x_s\right)\right),l_s^{\delta}\right)\\
&+E_{\theta,\xi}\sum_{s<t}\sum_{y}q\left(\xi\left(x_s\right)\right)\left(y\right)\log\frac{l_s^{\delta}\left(y\right)}{l^k\left(d_s,x_s\right)\left(y\right)}\\
&\leq 2\left(\frac{\Delta^k}{\delta}\log\frac{1}{\delta}+6\delta\right)t+k^2\log\frac{1}{\delta}+k\left(d_t^1+d_t^2\right).
\end{aligned}$$

Take $\delta=\max\left(\sqrt{\Delta^k},\frac{1}{k}\right)$ to finish the proof of the Proposition.

C.2. **Convergence $l_\omega\to q$.**

Proposition 6. *Suppose that instance process $x_1,x_2,...$ satisfies the sufficient data condition. Then,*

$$\lim_{t\to\infty}\frac{1}{t}E_{\theta,\xi}\sum_{s<t}\left\|q\left(\xi\left(x_s\right)\right)-l_\omega\left(d_s,x_s\right)\right\|=0. \tag{C.2}$$

Denote behavioral rules (see Section 5.4)

$$a^k = a_{\max} \circ l^k \text{ and } a_\omega = a_{\max} \circ l_\omega.$$

By Proposition 5 and convergence (B.2),

$$\lim_{k\to\infty} U\left(a^k;\omega,\bar{x}\right) = \lim \inf_{t\to\infty} E_{\xi,\theta}\frac{1}{t}\sum_{s<t} u\left(q\left(\xi\left(x_s\right)\right),\theta\left(x_t\right)\right).$$

Let $E_{\xi|d}$ denote the expectation with respect to realization of the auxiliary variable ξ, conditional on the observed database of past cases d_t. Notice that

$$E_{\xi|d} q\left(\xi\left(x_s\right)\right)(y) = \omega\left(\theta\left(x_s\right)|d_s\right) = l_\omega\left(d_s,x_s\right).$$

By standard convexity arguments, this implies that

$$U\left(a_\omega;\omega,\bar{x}\right) \leq \lim \inf_{t\to\infty} E_{\xi,\theta}\frac{1}{t}\sum_{s<t} u\left(q\left(\xi\left(x_s\right)\right),\theta\left(x_t\right)\right)$$

with equality if and only if (C.2) holds. Because for each k

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

Proposition 5 together with (B.2) imply that (C.2) must be true.

C.3. **Proof of Proposition 3.** The Proposition is a consequence of Propositions 5, 6, and convergence (B.2).

C.4. **Proof of Theorem 1.** Due to Proposition 6, it is enough to show that

$$\lim \sup_{t\to\infty} \frac{1}{t} E_{\theta,\xi} \sum\nolimits_{s<t} \left\| l^{K_s}\left(d_s\left(K_s\right),x_s\right) - q\left(\xi\left(x_s\right)\right)\right\| = 0.$$

Because $D\left(p,p'\right)$ is continuous in $p,p' \in \Delta\left\{0,1\right\}$ and $D\left(p,p'\right)=0$ if and only if $p=p'$, it is enough to show that

$$\begin{aligned} 0 &= \lim \sup_{t\to\infty} \frac{1}{t} E_{\theta,\xi} \sum\nolimits_{s<t} D\left(q\left(\xi\left(x_s\right)\right), l^{K_s}\left(d_s\left(K_s\right),x_s\right)\right) \\ &= \lim \sup_{t\to\infty} \sum_k \frac{m_t^k}{t} E_{\theta,\xi} \frac{1}{m_t^k} \sum_{s<t:K_s=k} D\left(q\left(\xi\left(x_s\right)\right), l^{k}\left(d_s\left(K_s\right),x_s\right)\right) \end{aligned} \tag{C.3}$$

where

$$m_t^k = \#\left\{s<t : K_s = k\right\}.$$

Denote the last period in which k-assignment is used before period t

$$s_t^k = \max\left\{s<t : K_s = k\right\}$$

and let $s_t^k = \infty$ if k assignment is never used before period t. Then, by Proposition 5, (C.3) is not higher than

$$\limsup_{t\to\infty} \frac{1}{t} E_{\theta,\xi} \sum\nolimits_{s<t} D\left(q\left(x_s\right), l^{K_s}\left(d_s\left(K_s\right), x_s\right)\right)$$

$$\leq \limsup_{t\to\infty} 4 \sum_k \sqrt{\Delta^k} \left(\log \frac{1}{\Delta^k} + 6\right) \frac{m_t^k}{t} \tag{C.4}$$

$$+ \limsup_{t\to\infty} \frac{1}{t} \sum_{k:s_t^k<\infty} \left[k^2 \log k + k\left(d^1_{s_t^k} + d^2_{s_t^k}\right)\right] \tag{C.5}$$

By (B.2),

$$\lim_{k\to\infty} \sqrt{\Delta^k} = 0 \text{ and } \lim_{k\to\infty} \sqrt{\Delta^k} \log \frac{1}{\Delta^k} = 0.$$

Because the process satisfies the sufficient data condition, $\lim_{t\to\infty} K_t = \infty$ and, for each k, $\lim_{t\to\infty} \frac{m_t^k}{t} = 0$. Because $\sum_k \frac{m_t^k}{t} = 1$, term (C.4) of the inequality above is equal to 0.

By the definition of d^i,

$$d_t^1 + d_t^2 \geq \sqrt{t}.$$

By the definition of switching rule K_t in equation (5.1), if $s_t^k < \infty$, then,

$$d^1_{s_t^k} + d^2_{s_t^k} \leq \frac{s_t^k}{k^3} \text{ and}$$

$$\sqrt{s_t^k} \leq d^1_{s_t^k} + d^2_{s_t^k} \leq \frac{s_t^k}{k^3}.$$

Therefore, if $s_t^k < \infty$, then

$$k \leq \sqrt[6]{s_t^k} \leq \sqrt[6]{t} := k_t^{\max}.$$

This permits to bound term (C.5). Because $\log k \leq k$,

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{k:s_t^k<\infty} \left[k^2 \log k + k\left(d^1_{s_t^k} + d^2_{s_t^k}\right)\right]$$

$$\leq \limsup_{t\to\infty} \frac{1}{t} \sum_{k=1}^{k_t^{\max}} k^3 + \limsup_{t\to\infty} \frac{1}{t} \sum_{k=1}^{K_t} k\left(d^1_{s_t^k} + d^2_{s_t^k}\right) + \limsup_{t\to\infty} \frac{1}{t} \sum_{k=K_t+1}^{k_t^{\max}} k \frac{s_t^k}{k^3}$$

$$\leq \lim_{t\to\infty} \frac{1}{t} \left(k_t^{\max}\right)^4 + \limsup_{t\to\infty} \frac{K_t^2 \left(d_t^1 + d_t^2\right)}{t} + \limsup_{t\to\infty} \sum_{k=K_t+1}^{\infty} \frac{1}{k^2}$$

$$\leq \lim_{t\to\infty} t^{-\frac{1}{3}} + \limsup_{t\to\infty} \frac{1}{K_t^3} + \limsup_{t\to\infty} \frac{1}{K_t} = 0.$$

Appendix D. Proof of Proposition 4

Let $\Delta_I = \Delta\{0,1\}^I$ be the distribution of outcomes inside each category. Distribution ω can be stated as a distribution over product of C elements of Δ_I : $\omega \in \Delta(\Delta_I)^C$. Because of exchangeability of categories, ω is De Finetti exchangeable. By an extension of the De Finetti's Theorem to Polish spaces ((Hewitt and Savage 1955), (Kingman 1978)), there is a distribution $\mu_\omega \in \Delta(\Delta_I)$ measurable function $q : [0,1] \to \Delta_I$ such that, for any finite set of objects $O' \subseteq O$, any assignment $y : O' \to \{0,1\}$,

$$\begin{aligned}&\omega(\theta(o) = y(o) \text{ for } o \in O')\\ &= \int_{\Delta_I} d\mu_\omega(\omega_I) \prod_{c\in C} \omega_I(\theta(c,i) = y(c,i) \text{ for all } i \text{ st. } (c,i) \in O').\end{aligned}$$

Because of exchangeability inside categories, measure μ_ω assigns probability 1 to exchangeable measures $\omega_I \in \Delta_I = \Delta\{0,1\}^I$. By the De Finetti's theorem, any exchangeable measure ω_I is represented by a distribution $\mu_I \in \Delta[0,1]$ so that

$$\begin{aligned}&\omega(\theta(o) = y(o) \text{ for } o \in O')\\ &= \int_{\Delta_I} d\mu_\omega(\mu_I) \prod_{c\in C} \int_{[0,1]} d\mu_I(p) \prod_{i:(c,i)\in O'} p^{y(c,i)}(1-p)^{1-y(c,i)}.\end{aligned}$$

Then,

$$\begin{aligned}A &:= \omega(\theta(c_2,i_2) = \theta(c_0,i_0) = \theta(c_0,i_0'))\\ &= \int_{\Delta_I} \left(\begin{array}{c} \left[\int_{[0,1]} p^2 d\mu_I(p)\right]\left[\int_{[0,1]} p d\mu_I(p)\right] \\ + \left[\int_{[0,1]} (1-p)^2 d\mu_I(p)\right]\left[\int_{[0,1]} (1-p) d\mu_I(p)\right] \end{array} \right) d\mu_\omega(\mu_I)\\ &= \int_{\Delta_I} \left(\left[\int_{[0,1]} d\mu_I(p)\, p\right]^3 + \left[\int_{[0,1]} d\mu_I(p)(1-p)\right]^3 \right) d\mu_\omega(\mu_I)\\ &+ \int_{\Delta_I} \left(\mathrm{Var}_{\mu_I} p \left[\int_{[0,1]} p d\mu_I(p)\right] + \mathrm{Var}_{\mu_I} p \left[\int_{[0,1]} (1-p) d\mu_I(p)\right] \right) d\mu_\omega(\mu_I)\\ &= \int_{\Delta_I} \left(\left[\int_{[0,1]} d\mu_I(p)\, p\right]^3 + \left[\int_{[0,1]} d\mu_I(p)(1-p)\right]^3 \right) d\mu_\omega(\mu_I)\\ &+ \int_{\Delta_I} (\mathrm{Var}_{\mu_I} p)\, d\mu_\omega(\mu_I),\end{aligned}$$

where Var_{μ_I} is the variance of random variable p with respect to distribution μ_I. Similarly,

$$\begin{aligned}
B &:= \omega\left(\theta\left(c_0, i_0\right) = \theta\left(c_0, i_0'\right)\right) \\
&= \int_{\Delta_I} \left(\left[\int_{[0,1]} d\mu_I\left(p\right) p \right]^2 + \left(\int_{[0,1]} d\mu_I\left(p\right) \left(1-p\right) \right)^2 \right) d\mu_\omega\left(\mu_I\right) \\
&+ 2 \int_{\Delta_I} \left(\mathrm{Var}_{\mu_I} p\right) d\mu_\omega\left(\mu_I\right),
\end{aligned}$$

$$\begin{aligned}
C &:= \omega\left(\theta\left(c_2, i_2\right) = \theta\left(c_0, i_0\right) = \theta\left(c_1, i_1\right)\right) \\
&= \int_{\Delta_I} \left(\left[\int_{[0,1]} d\mu_I\left(p\right) p \right]^3 + \left[\int_{[0,1]} d\mu_I\left(p\right) \left(1-p\right) \right]^3 \right) d\mu_\omega\left(\mu_I\right),
\end{aligned}$$

$$\begin{aligned}
D &:= \omega\left(\theta\left(c_0, i_0\right) = \theta\left(c_1, i_1\right)\right) \\
&= \int_{\Delta_I} \left(\left[\int_{[0,1]} d\mu_I\left(p\right) p \right]^2 + \left(\int_{[0,1]} d\mu_I\left(p\right) \left(1-p\right) \right)^2 \right) d\mu_\omega\left(\mu_I\right).
\end{aligned}$$

Denote $x = \int_{\Delta_I} \left(\mathrm{Var}_{\mu_I} p\right) d\mu_\omega\left(\mu_I\right) \geq 0$. Then,

$$\begin{aligned}
\omega\left(\theta\left(c_2, i_2\right) = \theta\left(c_0, i_0\right) | \theta\left(c_0, i_0\right) = \theta\left(c_0, i_0'\right)\right) &= \frac{A}{B} = \frac{C+x}{D+2x}, \\
\omega\left(\theta\left(c_2, i_2\right) = \theta\left(c_0, i_0\right) | \theta\left(c_0, i_0\right) = \theta\left(c_1, i_1\right)\right) &= \frac{C}{D}.
\end{aligned}$$

Finally, observe that

$$\begin{aligned}
&2C - D \\
&= 2 \int_{\Delta_I} \left(1 - 3 \left[\int_{[0,1]} d\mu_I\left(p\right) p \right] + 3 \left[\int_{[0,1]} d\mu_I\left(p\right) p \right]^2 \right) d\mu_\omega\left(\mu_I\right) \\
&- \int_{\Delta_I} \left(1 - 2 \left[\int_{[0,1]} d\mu_I\left(p\right) p \right] + 2 \left[\int_{[0,1]} d\mu_I\left(p\right) p \right]^2 \right) d\mu_\omega\left(\mu_I\right) \\
&= \int_{\Delta_I} \left(1 - 4 \left[\int_{[0,1]} d\mu_I\left(p\right) p \right] + 4 \left[\int_{[0,1]} d\mu_I\left(p\right) p \right]^2 \right) d\mu_\omega\left(\mu_I\right) \\
&= \int_{\Delta_I} \left(1 - 2 \left[\int_{[0,1]} d\mu_I\left(p\right) p \right] \right)^2 d\mu_\omega\left(\mu_I\right) \geq 0.
\end{aligned}$$

Hence, $\frac{C}{D} \geq \frac{1}{2}$ and

$$\frac{A}{B} = \frac{C+x}{D+2x} \leq \frac{C}{D}.$$

University of Chicago, Department of Economics
E-mail address: mpeski@uchicago.edu