Navarro, D. J. and Perfors, A. (2011). Hypothesis generation, the positive test strategy, and sparse categories. Psychological Review, 118, 120-34
http://dx.doi.org/10.1037/a0021110

Hypothesis generation, sparse categories and the positive test strategy

Danielle J. Navarro
School of Psychology
University of Adelaide

Amy F. Perfors
School of Psychology
University of Adelaide

Abstract

We consider the situation in which a reasoner must induce the rule that explains an observed set of data but the hypothesis space of possible rules is not explicitly enumerated or identified. The first part of the paper demonstrates that as long as hypotheses are sparse (i.e., index less than half of the possible entities in the domain) then a positive test strategy is near-optimal. The second part of this paper then demonstrates that a preference for sparse hypotheses (a sparsity bias) emerges as a natural consequence of the family resemblance principle; that is, it arises from the requirement that good rules index entities that are more similar to one another than they are to entities that do not satisfy the rule.

Introduction

Uncovering the rules that govern the observations that we make is a fundamental inductive inference problem, covering many substantively different domains and several formally distinct learning problems. In its most general form, the learner must induce the rule on the basis of a collection of observations and some information as to which observations satisfy the rule. Across domains, this broad problem includes children acquiring grammatical rules (e.g., Chomsky, 1957), scientists searching for physical laws (e.g., Popper, 1935/1990; Kuhn, 1970), people learning rule-governed concepts (e.g., Bruner, Goodnow, & Austin, 1956) and many others. In this paper we consider the problem of active learning, in which it is the learners responsibility to make queries regarding the phenomenon of interest in order to uncover the true rule that describes it. This learning problem has two distinct parts: the *hypothesis generation problem*, in which plausible candidate rules must be proposed, and the *hypothesis testing problem*, in which appropriate tests of those hypotheses must be constructed.

The hypothesis testing problem is well-studied in the literature on reasoning and decision-making, and displays a striking empirical regularity. In general, people prefer

to employ a *positive test strategy*, or PTS (see Nickerson, 1998; McKenzie, 2005, for an overview). The PTS can be characterized as the tendency to ask questions that will yield an affirmative response if the hypothesis currently under consideration is true (Klayman & Ha, 1987). Although it is sometimes difficult to disentangle from the matching bias (Evans, 1972; Evans & Lynch, 1973; Evans, 1998, see also Yama, 2001), the PTS is pervasive. It is observed in rule-learning problems (e.g, Wason, 1960; Taplin, 1975; Tweney et al., 1980; Klayman & Ha, 1989), the four-card selection task (e.g., Wason, 1968; Jones & Sugden, 2001), scientific research (e.g., Mahoney & de Monbruen, 1977; Mynatt, Doherty, & Tweney, 1978; Dunbar, 1993), and many other contexts (Nickerson, 1998). The bias to use a PTS can be ameliorated in some situations (e.g., Johnson-Laird, Legrenzi, & Legrenzi, 1972; Cheng & Holyoak, 1985; Cosmides, 1989), but is rarely completely eliminated. Moreover, although there are some senses in which it represents a logical error (Wason, 1960; Platt, 1964; Wason, 1968; Johnson-Laird & Wason, 1970), the PTS can be a highly effective learning strategy when certain assumptions are met (Klayman & Ha, 1987, 1989; Oaksford & Chater, 1994; Austerweil & Griffiths, 2008).

Hypothesis generation is less well-studied by comparison, but is presumably tied to the question of what kinds of rules people find to be *a priori* more plausible than others. Within the rule-based categorization literature, for instance, it is typical to assume the existence of a relevant class of possible rules (Nosofsky, Palmeri, & McKinley, 1994; Ashby & Gott, 1988; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Erickson & Kruschke, 1998), and there has long been a recognition that learning involves strategic shifts in the learner's choice of hypothesis (e.g. Goodnow & Pettigrew, 1955; Levine, 1959; Brown, 1974). There has been some exploration of the hypothesis generation problem in more general contexts than categorization (see Gettys & Fisher, 1979; Gettys, Mehle, & Fisher, 1986; Koehler, 1994; Thomas, Dougherty, Sprenger, & Harbison, 2008), usually in isolation from the hypothesis testing problem. Those studies that have linked the generation problem to the testing problem have tended to focus on empirical questions about how people's strategies for hypothesis formation affect their overall performance (e.g., Klahr & Dunbar, 1988; Farris & Revlin, 1989; Adsit & London, 1997). There is relatively little work addressing the theoretical question of which hypothesis formation strategies *should* be followed; the closest work that we are aware of is that of Cherubini, Castelvecchio, and Cherubini (2005), who present empirical evidence suggesting that people incrementally form hypotheses that are informative with respect to observed data.

In this paper we present an analysis of the active learning problem that considers the hypothesis testing and hypothesis generation problems together. In the first part of the paper, we extend previous results that derive the PTS as a near-optimal learning strategy when hypotheses are *sparse* (that is, when they index only a minority of possible entities). In the second, more novel part of the paper, we turn to the hypothesis generation problem and show that sparsity itself is a consequence of the more basic requirement that rules correspond to "good" categories. We conclude by discussing some of the assumptions and simplifications incorporated into this analysis.

Positive Tests for Sparse Hypotheses

Several studies suggest that the positive test strategy can be a rational learning strategy (Oaksford & Chater, 1994; Austerweil & Griffiths, 2008; Klayman & Ha, 1987) as long

as one key assumption is met. All of the formal results invoke some form of "sparsity assumption", in which the hypothesis is consistent with only a minority of logically possible observations (see McKenzie, 2005, for a discussion). Although the psychological ideas behind these results are quite general, they are limited in two respects, both of which Klayman and Ha (1987) briefly discussed but did not include in their formal analysis.

First, the derivations all assume that the learner considers only a single hypothesis at a time. Although there is evidence that people often do this (e.g., Taplin, 1975; Mynatt, Doherty, & Tweney, 1977; Doherty, Mynatt, Tweney, & Schiavo, 1979), it is not always the case. Many studies explore how entertaining multiple hypotheses affects the process of inference, with some finding that it helps (e.g., Klahr & Dunbar, 1988; Klayman & Ha, 1989), some finding that it does not (e.g., Tweney et al., 1980; Freedman, 1992), and some finding that it depends (e.g., Farris & Revlin, 1989; McDonald, 1990). In view of this, it would be useful to extend the existing formal results to cover the multiple hypothesis case.

Second, previous derivations analyze the PTS as a method for uncovering as much information as possible about the learner's hypothesis, not as a method for optimally identifying the true rule. While these are closely related goals, they are not necessarily equivalent. If the goal of the learner is to identify the true rule rather than simply to test the current hypothesis, it is important to have a formal analysis that explores the effect of different learning strategies on *all* of the candidates in the learner's hypothesis space – on all of the possible rules that they could consider.

In this section we retain the critical assumption of sparsity, but extend previous results by addressing the two issues discussed above. That is, we present an analysis that accommodates multiple hypotheses, and derives the PTS in a situation where the goal is to uncover the correct rule as quickly as possible (rather than extract information about the current hypothesis). Our results suggest that the effectiveness of the PTS can change as hypotheses are eliminated and the proportion of hypotheses that are explicitly available to the learner changes.

The ideal learner case

The kind of active learning problems in which we are interested is best exemplified by the traditional game of "twenty questions." In this game, one player (the oracle) thinks of an object, and the other player (the learner) can pose queries to the oracle. These queries must be yes-or-no questions, and the oracle must answer truthfully. Strictly speaking, the learner's goal in this game is to ask questions in such a way as to identify the object using 20 questions or fewer, but in practice the goal is to do so using as few questions as possible. An interesting variation of the game is the "rule learning" task introduced by Wason (1960). The rule learning game differs from twenty questions in that the oracle thinks of a rule rather than an object, and constrains the allowable queries to be of the form "does x satisfy the rule?" where x is an object. For example, in one game the oracle might think of a rule about numbers, such as PERFECT SQUARES. The learner's queries might include items like 7, 15, and 16, to which the oracle would reply "no", "no" and "yes" respectively. In the analysis below, we assume that the learning task is the Wason variant.

How would an ideal learner approach this game? Following the approach taken in the formal literature on rule-based categories (Nosofsky et al., 1994; Ashby & Gott, 1988; Goodman et al., 2008; Erickson & Kruschke, 1998) we assume that the learner has a set

of (not necessarily explicit) plausible hypotheses $\mathcal{H}$ about the rule. This is the learner's hypothesis space. (We defer questions about the origin of $\mathcal{H}$ to the next section.) Let h be a specific hypothesis about a rule, x be one possible query item, and $\mathcal{X}$ be the set of all possible queries that the learner might ask in the game. For simplicity, we suppose that the learner places an initial uniform prior over the hypotheses (so that $P(h) = 1/N$, where N is the current total number of hypotheses in $\mathcal{H}$). If the learner poses query x and the oracle gives response r, Bayes' theorem tells us that the degree of belief associated with hypothesis h is given by:

$$P(h \mid r, x) = \frac{P(r \mid h, x)P(h)}{\sum_{h'} P(r \mid h', x)P(h')}. \tag{1}$$

The term $P(r \mid h, x)$ denotes the probability that the oracle would have given response r to query x if hypothesis h were true. This probability is 1 if the query item is contained in the rule corresponding to the hypothesis and 0 if it is not. Thus, all hypotheses that are inconsistent with the oracle's answer are eliminated and all others are retained. If we let n_c denote the number of hypotheses that are consistent with the set of responses and queries so far, then the degree of belief associated with hypothesis h is now given by:

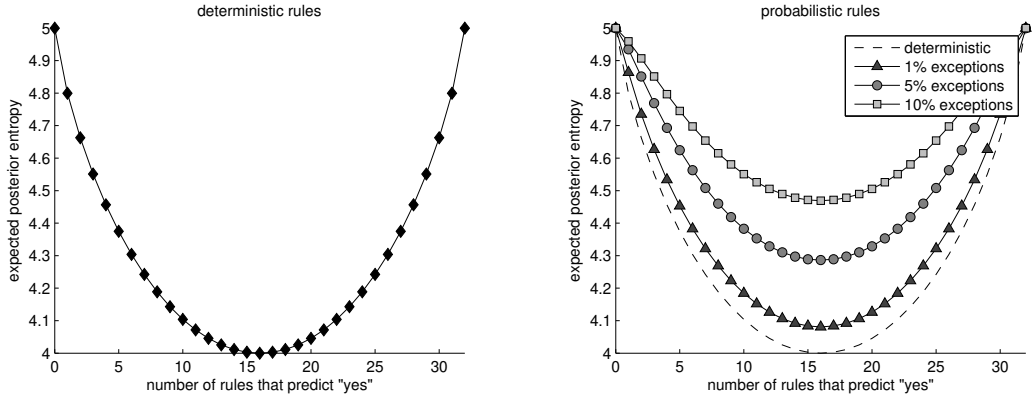
$$P(h \mid r, x) = \begin{cases} \frac{1}{n_c} & \text{if } h \text{ is consistent with all responses} \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Because any hypothesis that is not consistent with the responses so far is eliminated, at every stage in the game the learner has a collection of not-yet-falsified hypotheses, each of which is equally plausible. Though simple, this setup closely mirrors the framework used by Popper (1935/1990), in which a single disconfirming observation is sufficient to falsify a hypothesis, but many confirming observations are needed to provide strong evidence in favor of the hypothesis. It is consistent with the classic "win-stay/lose-shift" strategy pervasive in human learning (Goodnow & Pettigrew, 1955; Levine, 1959; Brown, 1974), in which people retain a hypothesis as long as it makes correct predictions, but discard it the moment it makes an incorrect one.

The learner's goal is to choose a query x in such a way as to allow the true rule to be identified as quickly as possible. The learner knows that the oracle is either going to say "yes", in which case there will be some number of hypotheses n_y remaining, or the oracle is going to say "no", in which case there will be $n_n = N - n_y$ hypotheses remaining. For instance, one query might be consistent with only 1% of possible rules, while another one could be consistent with 50% of the rules. For the first query, there is a 1% chance that the oracle will eliminate 99% of the possible rules and a 99% chance that it will eliminate 1% of the rules; for the second, 50% of the rules will end up eliminated regardless of what the oracle says.

From an information theoretic perspective (e.g., MacKay, 2003) it is not difficult to show that the second type of query is superior. If the aim is to identify the rule as quickly as possible, a rational learner should choose the item x that is expected to minimize the posterior entropy $H(h \mid x, r)$ of her belief about the identity of the correct rule, since this corresponds to a state of maximal knowledge and minimal uncertainty.[1] The learner should

[1]There are, of course, other ways that the learner's goal could be formalized (see, e.g., Nelson, 2005,



(a) (b)

Figure 1. Expected posterior entropy (in bits) as a function of n_y, for a domain with $N = 32$ objects. Panel a shows the result for deterministic rules (Equation 3), whereas panel b shows what happens when some proportion of the items ϕ are allowed to constitute exceptions to the rules. The qualitative pattern is the same in all cases, regardless of whether the rules are deterministic or probabilistic.

therefore pick the x that is expected to return the most information about the true rule. Formally, the expected entropy is given by:

$$E_r[H(h \mid x, r)] \quad = \quad \frac{n_y \ln n_y + n_n \ln n_n}{N}, \tag{3}$$

where $n_y + n_n = N$. The important thing to note is that this function, which is explicitly derived in the Appendix and illustrated for $N = 32$ in Figure 1a, is minimized when $n_y = n_n = N/2$. Thus, the optimal query x is one that is true for exactly half of the not-yet-eliminated hypotheses. This corresponds to the "bisection" search method that people intuitively prefer to use in simpler situations. Games in which the learner needs to identify an unknown person often start with queries about gender for exactly this reason.

Extension to probabilistic rules

The formalism for active learning described earlier assumes that rules are deterministic, allowing no exceptions, and that the oracle gives perfect feedback. This is of course a simplification, given that many (perhaps most) rules allow exceptions. We can capture this by supposing that the proportion of exception items is ϕ (or, equivalently, that the oracle makes mistakes for a proportion ϕ of the items). When this occurs, no rule is ever perfectly falsified, since any disagreements between the rule and the oracle's answer might have been errors by the oracle or genuine exceptions to the rule. Under such circumstances, the calculation of the expected posterior entropy is somewhat more complicated: a derivation appropriate to this case is provided in the Appendix. The important point, however, is that

for a discussion), including minimizing the number of queries or maximizing the diagnosticity of the next question. We chose this because it is a reasonable goal, and because it is consistent with empirical findings about how human learners appear to gather information (Nelson, Tenenbaum, & Movellan, 2001).

the qualitatively important features of the expected entropy function are unchanged. This is illustrated in Figure 1b, which plots the expected posterior entropy as a function of n_y for several different levels of ϕ. The key point is that the curves are all still U-shaped, with a minimum value at $n_y = N/2$. Since these characteristics are the ones that our analysis relies upon, it is clear that our results generalize to probabilistic rules.

The partial-information case

The previous analysis relies on the assumption that the learner has access to $\mathcal{H}$, the full set of plausible hypotheses, and is able to choose x in such a way as to ensure that $n_y = n_n$. Under these assumptions, there ought not to be any PTS bias. However, these assumptions are not satisfied in the Wason (1960) task or the twenty questions game more generally. In most cases, the learner has only very limited information to work with: for instance, it is typically assumed that the learner can only keep a small number of hypotheses in working memory, perhaps only one hypothesis (see Dougherty & Hunter, 2003). If so, it is likely that only this limited subset of explicit hypotheses $\mathcal{H}_E$ can be used to guide the choice of x. For the current purposes, we assume that $\mathcal{H}_E$ is a more-or-less random subset of $\mathcal{H}$ (e.g. Nahinsky, 1970; Williams, 1971). As long as $\mathcal{H}$ only contains "plausible hypotheses" this assumption seems reasonable: it amounts to the assumption that people ignore "implausible" hypotheses, and choose randomly among the "plausible" ones. As such it is broadly consistent with models of hypothesis generation (Gettys & Fisher, 1979; Thomas et al., 2008) which assume that people form hypotheses in a sensible fashion.

To formalize the hypothesis testing problem when the learner has access only to the explicit subset $\mathcal{H}_E$, note that the set of N rules and M entities produces an $N \times M$ binary "truth matrix", whose ijth element is 1 if the ith hypothesis in $\mathcal{H}$ predicts an affirmative response to the jth possible query, and 0 if it does not. In the simplest case, this is an unstructured matrix in which the cells are independent of one another. We let θ denote the probability that any element takes on value 1. In this situation, if the learner chooses x completely at random, then the number of hypotheses that predict an affirmative response to any given query will be binomially distributed:

$$n_y \sim \text{Binomial}(\theta, N). \tag{4}$$

If it happens to be the case that $\theta = 1/2$, then the expected number of hypotheses that would yield an affirmative response n_y is $N/2$; in other words, the query has a reasonable chance of being optimal with respect to $\mathcal{H}$. However, for other values of θ this is not the case. In particular, if the hypothesis space is *sparse*, then $\theta < 1/2$, meaning that most hypotheses would yield an affirmative response only to a minority of possible queries. In this situation most queries will be suboptimal, since n_y will probably be smaller than n_n. If the learner has no hypotheses in mind (i.e., $\mathcal{H}_E$ is empty) then there is nothing the learner can do to improve matters. But if she has a small number of hypotheses $\mathcal{H}_E$ in mind, then choosing queries to which those hypotheses yield affirmative responses will boost n_y, and thus improve the efficiency of the query. This is true even if the learner does not know the sparsity, and even if the explicit hypotheses $\mathcal{H}_E$ themselves are not sparse, as long as the average sparsity θ of the hypotheses in the entire space is less than $1/2$; the entropy is related to how efficiently the query eliminates hypotheses across the entire hypothesis space.

Summary

The basic result states that as long as the hypotheses tend to be sparse and the learner does not have access to all relevant hypotheses at all times, it is sensible to adopt the PTS with respect to the set of hypotheses that the learner *does* have access to. This is true whether the rules are probabilistic or deterministic. This occurs because a sparse hypothesis space means the oracle is expected to produce too many "no" responses with respect to $\mathcal{H}$, and a strategy that is highly biased towards a "yes" response with respect to $\mathcal{H}_E$ is the best way to overcome it. This does not mean that such a strategy will *necessarily* or *entirely* counteract such a bias – it depends on the strength of the bias and the details of the specific rules and hypothesis space. It does mean that, in general, the PTS will counteract it better than most other strategies would. This is because, even if some rules are not currently available to the learner, the oracle's response will still be informative about them (ruling out those that are inconsistent from ever being considered).

An interesting corollary of this result is the implication that as the number of implicit hypotheses decreases or the sparsity of the remaining hypotheses increases the extent of the bias should reduce. When all remaining hypotheses are explicit the bisection strategy will become optimal. In fact, empirical findings do suggest that beginning with a confirmatory strategy and moving toward a disconfirmatory one is more effective (Mynatt et al., 1978). Our analysis can help to explain other experimental results as well. For instance, the notion that the PTS may emerge because of a capacity limitation is consistent with empirical evidence that increasing the number of alternative hypotheses considered in $\mathcal{H}_E$ may improve learning (Klayman & Ha, 1989).

Hypothesis Generation, Sparse Categories and Family Resemblances

The main assumption made by all rational analyses of the PTS is that hypotheses are sparse. Regardless of the precise set up, some assumption of this form appears to be necessary. In light of this, it is natural to ask what theoretical justification exists to support this assumption. While there are some situations in which hypotheses are necessarily sparse (Austerweil & Griffiths, 2008), most analyses have relied on the fact that the sparsity assumption appears to be *empirically* justified (Klayman & Ha, 1987; Oaksford & Chater, 1994). While this is reassuring, it is somewhat unsatisfying at a theoretical level: a theoretical account of the PTS that makes use of the sparsity assumption should, ideally, be accompanied by a theoretical explanation of sparsity itself. In this section, we seek to provide this theoretical basis for sparsity. We begin by discussing the evidence for the sparsity assumption, and go on to derive sparsity from more basic premises.

The empirical evidence for sparsity

Regardless of what the correct theoretical explanation for the phenomenon is, there is considerable empirical evidence that people do have sparse hypothesis spaces. Most obviously, people's hypotheses about rules are presumably heavily reliant on natural categories. For instance, in Wason's number game, the learner might propose a sparse rule such as "multiples of two" because he or she is relying on the natural category of "even numbers".

Therefore if natural categories are sparse, one would expect that the hypotheses that people form in the context of an active learning problem would also be sparse. It is intuitively obvious that natural categories are indeed sparse: only a minority of entities in the world belong to the category "dog", for instance. Even when we consider the more realistic situation where the domain is somewhat restricted, sparsity is still the norm. That is, even if we restrict ourselves to a discussion of animals (rather than all entities in the world) it is still the case that most animals are not dogs. Not surprisingly, therefore, experiments aimed at eliciting information about the structure of natural categories support the intuition that natural categories are sparse (e.g., De Deyne et al., 2008).

Even in the more specific context of hypothesis generation – as opposed to category learning generally – there is empirical evidence for sparsity. In previous work (Perfors & Navarro, 2009) we found that the number rules people spontaneously construct are extremely sparse. Similarly, in a more complex study in an automotive context (Mehle, 1982), participants appeared to prefer sparse theories; indeed, the more they knew about the domain, the sparser their hypotheses tended to be. In short, hypothesis sparsity is an empirical regularity regarding human hypothesis generation.

Sparsity is not a logical necessity

Just as it is obvious that natural categories are sparse, it is also obvious that this sparsity is not a logical necessity. If the domain in question consists of a set of M objects, then we can associate every possible rule h with the category of objects that satisfy the rule. In total, there are 2^M distinct categories that are possible in this domain, any of which might correspond to the category in question. Clearly, if the learner treats each of these categories as equally plausible, then no sparsity bias will emerge: for every category containing K entities, there exists a complementary category containing $M - K$ entities. As a consequence, the average sparsity of will be exactly $1/2$. While this is hardly a novel insight, it illustrates a simple but important point: sparsity requires a *psychological* explanation.

A natural place to look for such an explanation is to examine how sparsity is captured within formal models of human categorization (e.g., Nosofsky, 1984; Anderson, 1991; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Griffiths, Sanborn, Canini, & Navarro, 2008). Most of these models do possess a sparsity bias, and in the majority of cases this bias is imposed because the model assumes that categories form a *partition* of the stimuli. That is, the models assume that individual entities are assigned to exactly one category, and as a consequence it is almost impossible for the models to produce non-sparse category assignments. This is illustrated in Figure 2a, in which a partition of six objects has one category of size 3 `(A,B,C)`, one of size 2 `(D,F)` and one of size 3 `(E)`, corresponding to a very sparse representation. The key point here is that theoretical models tend to impose sparsity via the structure of the category representation, which in this instance is a partition.

Of course, it is not true in general that categories are organized into partitions. In fact, there is considerable evidence that natural concepts can be organized into a range of different structures: besides partitions, human mental representations can take the form of hierarchies, spaces, networks, grammars and many other possibilities (see Kemp & Tenenbaum, 2008, for an overview). Even so, it does appear to be the case that sparsity holds across these structures. To give a single example, the most common alternative to partitions

are hierarchies. As illustrated in Figure 2b, if every node in the hierarchy maps to a category, then the overall representation tends to be sparse. The only way to avoid sparsity is to exclude the "singleton" categories category (i.e., the terminal nodes), and then construct the most "top-heavy" tree possible: over 6 objects, this would produce the categories `(A,B)`, `(A,B,C)`, `(A,B,C,D)`, `(A,B,C,D,E)`, and `(A,B,C,D,E,F)` which has sparsity just over 1/2. However, this is an extremely atypical case: in general, trees tend to be sparse.

In view of this, one might be tempted to think formal models *necessarily* impose sparsity. However, this is not the case. For instance, in the connectionist approach to semantic representation the main commitment is to some form of "distributed representation" (Rogers & McClelland, 2004), which may or may not be sparse. Similarly, overlapping clustering (Shepard & Arabie, 1979) produces stimulus representations in which objects can belong to any number of categories, and there are no constraints on how categories overlap. Again, these representations can be sparse, but they do not have to be. For instance, Figure 2c shows a distributed representation over six objects produced by a set of overlapping nonsparse categories. Thus, while sparsity is common among the representational systems used in cognitive science, it is not a required feature of formal models in general. It is possible to devise representational systems that systematically violate sparsity,[2] but since this would be inconsistent with the empirical evidence, researchers have avoided doing so.

The key thing to take from the previous discussion is this: the general tendency to see sparse formalisms for category representation is a *consequence* of the empirical data, not an explanation of that data. To see this, suppose that we were to argue that – although sparsity is not a logical requirement – it emerges because people rely on structured representations (e.g., partitions) that are sparse. This argument would be correct as far as it goes, but it begs the question as to why some (sparse) formalisms are plausible, but other (non-sparse) possibilities are not. If we try to justify the preference for sparse formalisms by reference to the empirical data (human categories are sparse) then we are right back where we started: "explaining" sparsity by pointing out that human categories are sparse. In short, we have *no* explanation at all for the sparsity of categories. A more general theoretical principle is required.

The family resemblance principle as a potential explanation

At this point it is clear that we are looking for a psychological principle that is (a) satisfied by natural categories, (b) consistent with existing formal models, and (c) explains the emergence of the hypothesis sparsity effect. One candidate for this principle is *family resemblance*. Put simply, the family resemblance principle states that a category is a good

[2]Indeed, this is trivially easy to do. Let F denote some formal system that generates category system c (defined over a set of M objects) with probability p, and produces an expected sparsity of θ. Then, we may define F^* to be the formal system that generates the "complementary" category system c^* with probability p. By complementary we mean that if the k-th category in c indexes the set of q items $c_{k1}, \ldots, c_{kq}$, then the k-th category of c^* indexes the other $M-q$ items in the domain (as an example, the category systems shown in Figures 2a and 2c are complementary). The expected sparsity of representations produced by formal system F^* will necessarily be $1-\theta$. Thus, if F is sparse, then F^* is not. In exactly the same way that categories are not logically constrained to be sparse, neither are formal systems for organizing categories. As before, one might argue that the complementary system F^* to a sensible sparse system F tends to be somewhat implausible, but such an argument would rely on the very thing we are trying to explain: human preferences for sparse representations.

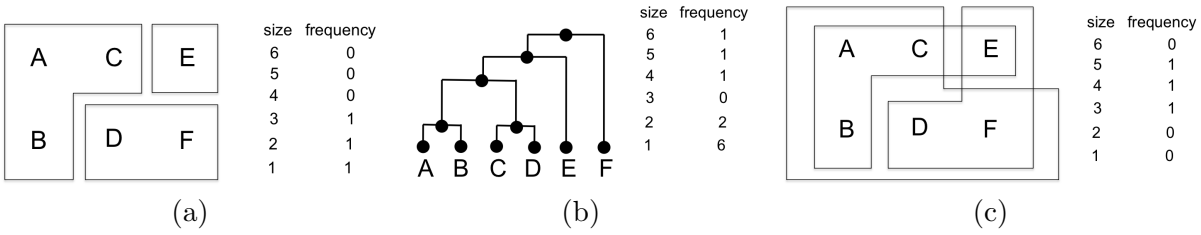


Figure 2. Are formal representations sparse? Partitions (panel a) are almost always sparse, trees (panel b) are usually sparse, while distributed representations (panel c) can be sparse or non-sparse as desired.

one if the members of the category tend to more similar to one another than they are to entities that do not belong to category. The fact that natural categories satisfy the family resemblance principle was pointed out by Rosch (1978), and the principle is reflected in some form in most models of categorization (e.g., Nosofsky, 1984; Anderson, 1991; Kruschke, 1992; Love et al., 2004; Griffiths et al., 2008). So it is clear that family resemblance meets the criteria (a) and (b) above. Additionally, variations of this principle provide much of the foundation for the statistical theory of classification (see Gordon, 1999), so there is something of a "rational flavor" to the idea. In short, the family resemblance principle is a central element in any psychologically plausible theory of concepts.

In the next section, we address the third criterion, and show that sparsity is a logical consequence of the family resemblance principle itself. When doing so, it is helpful to distinguish the "pure" idea of family resemblance from the formal models that implement this principle. As noted above, formal models often enforce sparsity by assuming that categories form a partition or a hierarchy, or some related structure that is necessarily sparse. However, these restrictions are not strictly required by the basic idea of family resemblance. The critical constraint that the family resemblance idea implies is this: a good category is one that groups together items that are more similar to one another than they are to other items. Thus the derivation that follows seeks to show that that sparsity follows from the "group by similarity" idea.

On the goodness of categories

To begin with, we need to formalize the family resemblance principle. Suppose that we have some hypothesized rule h that picks out a category, and let $\boldsymbol{x}_h$ denote the set of entities that belong to that category. If we let $s(x_i, x_j)$ denote the similarity between entities x_i and x_j, then the average "within category" similarity $s_{\text{in}}(\boldsymbol{x}_h)$ is given by

$$s_{\text{in}}(\boldsymbol{x}_h) = \frac{1}{m_h(m_h-1)} \sum_{x_i \in \boldsymbol{x}_h} \sum_{x_j \in \boldsymbol{x}_h} s(x_i, x_j) \tag{5}$$

where m_h denotes the number of entities in the category (out of a total of M entities in the domain), and the summations are taken over all $i \neq j$. This is graphically illustrated in Figure 3: in this figure, the darker shaded cells (those labelled "within") correspond to

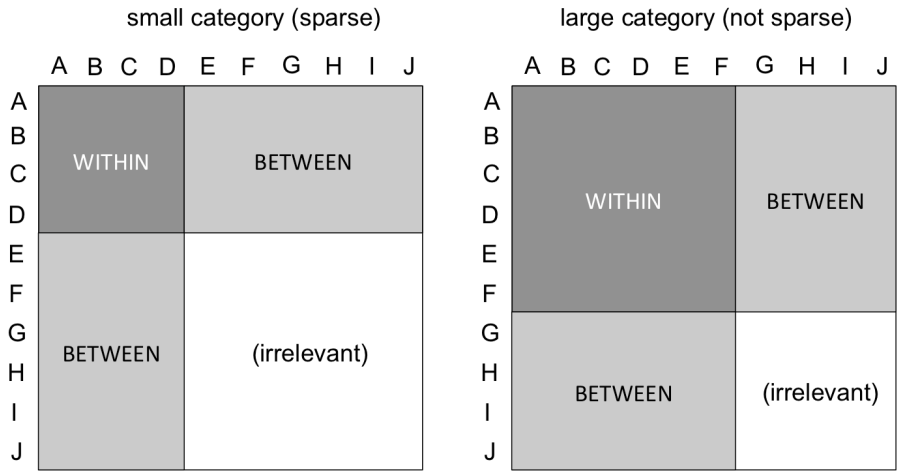


Figure 3. An illustration of which similarities contribute to the calculation of category goodness, for domain with $M = 10$ items. On the left, we have a small category containing 4 items (A,B,C,D) and on the right the category is large, with 6 items (A,B,C,D,E,F). In both figures, the dark shaded area corresponds to elements of the similarity matrix that contribute to the calculation of the within-category similarity, whereas the lightly shaded areas indicate which pairwise similarities contribute to the between-category similarity. Unshaded areas make no contribution, and are hence irrelevant. The key observation is that larger categories implicate a much larger proportion of the matrix.

similarities that are averaged in Equation 5. Following the same logic, the average similarity between category members and non-members, $s_{\text{out}}(\boldsymbol{x}_h)$ is

$$s_{\text{out}}(\boldsymbol{x}_h) = \frac{1}{2m_h(M - m_h)} \sum_{x_i \in \boldsymbol{x}^{(h)}} \sum_{x_j \notin \boldsymbol{x}^{(h)}} s(x_i, x_j). \tag{6}$$

By convention, this is generally referred to as the "between category" similarity, though in the current context the term is slightly odd since there is only a single contrast category (i.e., non-members). In any case, the relevant pairwise similarities for this expression correspond to the lighter shaded cells in Figure 3. Having defined the within- and between-category similarities, it is straightforward to formalize the family resemblance principle: the goodness of the category g_h is simply the difference between these two averages,

$$g_h = s_{\text{in}}(\boldsymbol{x}_h) - s_{\text{out}}(\boldsymbol{x}_h) \tag{7}$$

If $g_h > 0$ then the entities within the category are more similar to each other than they are to other entities, and the category is a good one in the family resemblance sense. In contrast, if $g_h < 0$, then the category members are actually more similar to non-members than they are to one another, and it is therefore a bad family resemblance category.

In what ways can the goodness of categories vary? Clearly, this depends in part on the precise nature of the pairwise similarity function $s(x_i, x_j)$. As such, there would be expected to be idiosyncratic variations across different stimulus domains. For instance, if similarities are constrained by a geometric structure (e.g. Attneave, 1950), we might expect

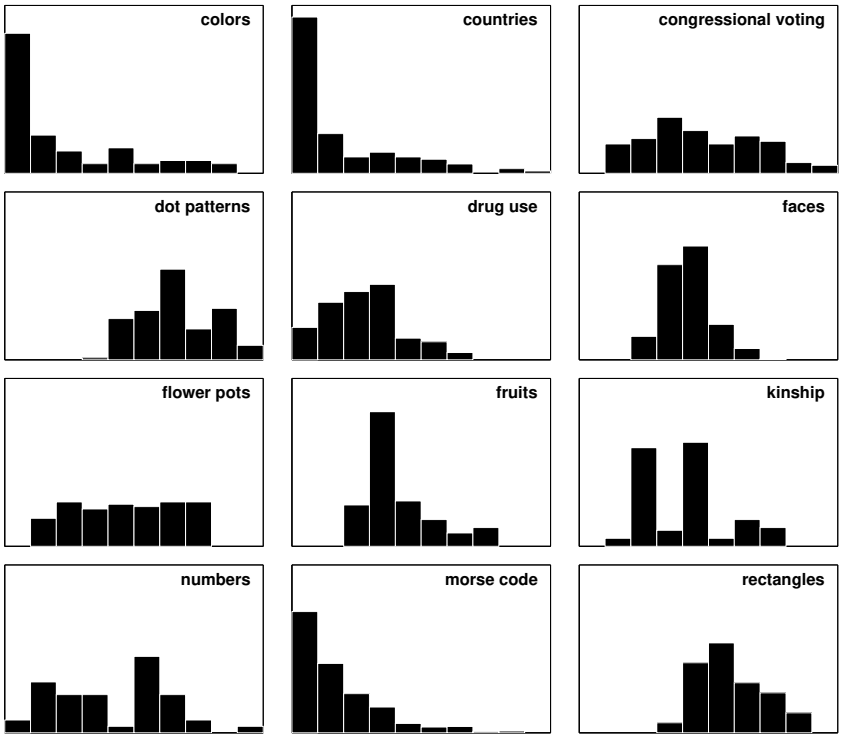


Figure 4. The marginal distribution over empirical similarities for several domains. All plots are drawn on the same scale, with the horizontal axis running from 0 to 1 since all data sets have been normalized so that similarities fall within this range. Across 12 data sets it is clear that similarities can be distributed in many different ways. The data sets are all available online courtesy of Michael Lee (`www.socsci.uci.edu/~mdlee/sda.html`) and cover a range of collection methodologies and domains. Specifically the data relate to colors (Ekman, 1954), countries (Navarro & Lee, 2002), congressional voting patterns (Romesburg, 1984), dot patterns (Glushko, 1975), patterns of drug use (Huba et al., 1981), photographs of faces (unpublished data, Michael Lee), drawings of flower pots (Gati & Tversky, 1982), fruits (Tversky & Hutchinson, 1986), kinship terms (Rosenberg & Kim, 1975), the numbers 0-9 (Shepard et al., 1975), morse code patterns (Rothkopf, 1957) and rectangles with interior line segments (Kruschke, 1993).

somewhat different answers than would be obtained if similarities are described in terms of feature overlap (Tversky, 1977). However, we wish to argue that while some things are specific, and depend on the particular structure that underlies the similarity function, other there are *also* some regularities that might be expected to be universal (or nearly so), and do not actually depend on the details of the similarity structure. In particular, we suggest that sparsity is one such universal.

With this in mind, in our derivation below we strip out almost everything that might plausibly be called "structure" from the similarity function. In effect, we treat the pairwise similarities as if they were independent and identically distributed random variables. Moreover, since the (marginal) distribution of empirically-measured similarities varies wildly across domains (see Figure 4) we make almost no assumptions about what distribution the pairwise similarities are generated from. The *only* constraint we impose is to assume that this distribution has finite mean and finite, non-zero variance. In other words, we assume that the pairwise similarities among entities are not all identical: there must be some variation in the world. The reason for taking this "minimalist" approach is as follows: our goal is to ensure that our derivation does not implicitly rely on any deeper representational structure, and thus show that sparsity follows from the family resemblance principle itself. The simplest way to do this is to remove any such structure. It should *not* be construed as a claim that no deeper structure exists in real life.

Given these extremely weak assumptions, what can we say abut the distribution of the category goodness measure g_h? While the raw distribution over *pairwise* similarities $s(x_i, x_j)$ can take any shape so long as it has finite mean μ and variance σ^2, the central limit theorem (e.g. Schervish, 1995, theorem B.97) implies that (in the limit) the within and between category *averages* become normally distributed

$$s_{\text{in}}(\boldsymbol{x}_h) \sim \text{Normal}\left(\mu, \frac{\sigma^2}{m_h(m_h-1)}\right) \tag{8}$$

$$s_{\text{out}}(\boldsymbol{x}_h) \sim \text{Normal}\left(\mu, \frac{\sigma^2}{2m_h(M-m_h)}\right). \tag{9}$$

However, the critical point here is not the fact that the distributions become normal, but rather the fact that the variance of these distributions depends on the number of items m_h that belong to the category. Taking the difference between these variables yields the following asymptotic distribution for the category goodness measure:

$$g_h \sim \text{Normal}\left(0, \frac{\sigma^2}{m_h(m_h-1)} + \frac{\sigma^2}{2m_h(M-m_h)}\right). \tag{10}$$

Notice that the value of σ just acts as a scaling constant and plays no role in determining the sparsity of the resulting hypothesis space. That is, if the combination σ_1, γ_1 yields an expected sparsity θ, then for every other possible variance σ_2 there exists a corresponding threshold γ_2 that yields the same sparsity θ. Conversely, if there is no value γ_1 that would yield sparsity θ if the variance is σ_1, then there is no value of γ_2 that would do so for σ_2. Thus, for simplicity and without loss of generality we set $\sigma = 1$, which gives

$$g_h \sim \text{Normal}\left(0, \frac{1}{m_h(m_h-1)} + \frac{1}{2m_h(M-m_h)}\right). \tag{11}$$

Now, notice that the function that describes the variance of the distribution over category goodness

$$\frac{1}{m_h(m_h-1)} + \frac{1}{2m_h(M-m_h)} \tag{12}$$

is largest when m_h is small.[3] The difference in variance means that it is "easier" to find a very good sparse rule than it is to find a very good non-sparse one. It also implies that it is easier to find a very bad sparse rule, but since the learner is presumably uninterested in finding bad rules, this is not particularly interesting.

This difference in variability emerges because sparser rules involve aggregating over fewer similarities. To understand why, consider Figure 3. It illustrates which cells in a 10×10 similarity matrix contribute to the category goodness calculation for a sparse category containing 4 objects (left) and non-sparse one containing 6 objects (right). The dark shaded area corresponds to elements of the similarity matrix that contribute to the calculation of the within-category similarity, whereas the lightly shaded areas indicate which pairwise similarities contribute to the between-category similarity. Unshaded areas make no contribution, and are hence irrelevant. Note that the rule indexing fewer items has many more unshaded cells, indicating that sparser rules involve fewer pairwise similarity calculations. What this means is that it is much easier to find a small collection of items that are especially similar to one another *or* especially dissimilar to one another. In essence, this is what the central limit theorem implies – the more pairwise similarities that we have to average over to compute the category goodness, the more likely it is that the category will be "average" (i.e., have g_h very close to zero).

The consequence of this is that the very best categories are sparse, as are the very worst ones. For instance, the numbers 2, 4 and 6 are much more similar to one another than most numbers, and so (2,4,6) is reasonably good category (i.e., $g_h > 0$). In contrast, 0, 2 and 7 are unusually dissimilar, and so (0,2,7) ends up being a bad category (i.e., $g_h < 0$). However, when one looks at the complementary categories (0,1,3,5,7,8,9) and (1,3,4,5,6,8,9), it is clear that these are *both* mediocre at best ($g_h \approx 0$). In the first case, there are some items that are very similar (the five odd numbers) to one another, but since these are lumped in with two non-odd numbers (0 and 8) it is not a very good category. In the second case, there is a run of consecutive numbers from 3 to 6, but there are three other numbers in the category, so it too is fairly poor. That is, small categories have the potential to be very good and very bad, while large categories do not.

This is not unique to the numbers domain, of course. The small category (mother, father, daughter, son) is very sparse, and feels like a very good category (it has $g_h \gg 0$) – it is put together from kinship terms that are unusually similar to one another. In contrast, the small category (grandmother, uncle, cousin, brother) feels like a very bad category (it has $g_h \ll 0$) – it is put together by choosing a deliberately weird grouping. The complementary categories in both cases consist of a large number of kinship terms, some of which are similar

[3] To be strictly accurate, this function is not perfectly monotonic. The minor violation of monotonicity is due to the fact that the category goodness calculation involves the difference between two averages (i.e., the within-category similarity and the between-category similarity), rather than a single average. As such, there is a minor violation of this effect that occurs when the rule in question is almost completely non-sparse (i.e., when $m_h \approx M$). However, as is clear from Figure 5 this weak violation of monotonicity does not induce a non-sparse hypothesis space.

and others are not, and hence have $g_h \approx 0$ in both cases.

Sparsity, finite memory and family resemblance

What does this tell us about sparsity? To answer this, recall that in a domain with M items, there are a total of 2^M distinct ways to group items together into a category. As the domain size becomes large, this number becomes so large that even a learner with a very large memory capacity will only be able to encode a small proportion of them. With this in mind, suppose that the learner retains only the best possible categories. Specifically, imagine that the learner has some threshold γ such that the rule h is an admissible member of $\mathcal{H}$ only if the goodness g_h is greater than that threshold. Notice that we place no particular *structural* constraints on which categories are included: for $\gamma > 0$ the learner ends up with a collection of good categories that need not form a tree, a partition or any other structure. Using the result in Equation 11, we can calculate the expected sparsity of the selected categories as a function of γ (see Appendix for technical details). This function is plotted in the left panel of Figure 5 for a domain consisting of $M = 16$ entities. As is clear from inspection, when the threshold is high enough ($\gamma > 0$) the hypothesis space tends to be sparse (that is, $\theta < 1/2$); and as the threshold gets higher, the average sparsity decreases.

One natural question to ask is whether this pattern occurs for real data sets, since the simplifying assumptions that we have made (such as independence) will not necessarily hold, and the analysis makes use of an asymptotic argument based on the central limit theorem. There are many empirical similarity matrices that have been published, and so it is straightforward to verify that the prediction holds. Examples constructed from three empirical similarity matrices are shown in the right panel of Figure 5. The numbers data is a similarity matrix that estimates the similarity between all pairs of numbers in the range 0-9 (Shepard et al., 1975). The kinship data (Rosenberg & Kim, 1975) examines the perceived similarity between 15 different kinship relations (e.g., brother, sister, cousin, father). Finally, the numbers data measures the perceived similarity between 16 nations (Navarro & Lee, 2002). In each case, we calculated the goodness for all 2^M possible categories (where $M = 10$ for numbers, $M = 15$ for kinship and $M = 16$ for countries), and then used these calculations to infer the sparsity θ that would emerge if the learner included only those categories with $g_h > \gamma$, for a range of γ values. As is clear from inspection, the empirical data show the expected pattern. That is, good family resemblance categories tend to be sparse.

We can see this in a little more detail if we calculate the category goodness for all possible categories in some domain, and plot this as a function of size. Figure 6 depicts the effects of three different choices of threshold γ given all $2^{15} = 32768$ logically possible categories for the kinship terms. The horizontal axis captures the number of entities m_h captured by the rule while the vertical axis captures the goodness of the rule g_h. As predicted, the sparse rules vary widely in goodness, whereas less sparse ones are all close to the average. As a result, when $\gamma = 0$, about half of the possible categories meet the threshold, and there is no sparsity bias in the corresponding hypothesis space $\mathcal{H}$ (on average, the rules admitted into $\mathcal{H}$ index 7.58 of the 15 items). However, when the threshold is raised (for instance, to $\gamma = 0.15$ or $\gamma = 0.075$), only the sparser rules are good enough to be included.

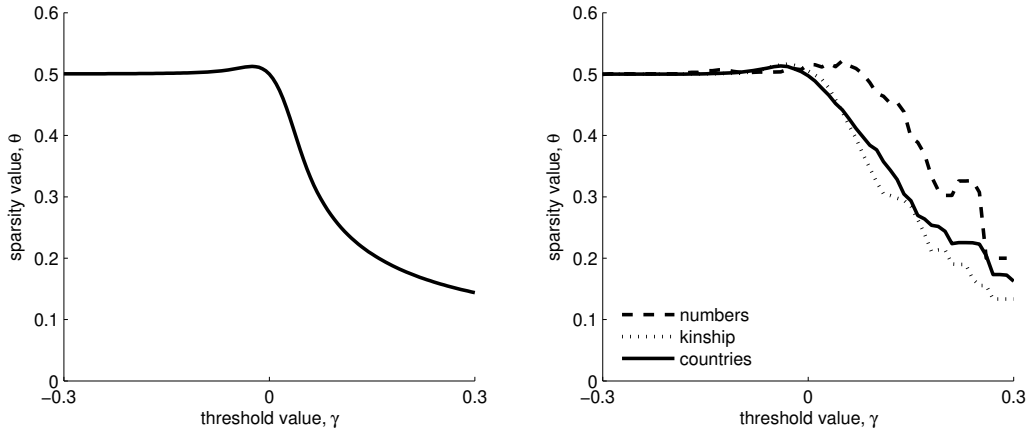


Figure 5. On the left, the expected sparsity θ of the hypothesis space $\mathcal{H}$ plotted as a function of the threshold γ, for a small domain with $M = 16$ entities. On the right, the corresponding functions for three empirical data sets, numbers ($M = 10$), kinship terms ($M = 15$) and countries ($M = 16$).

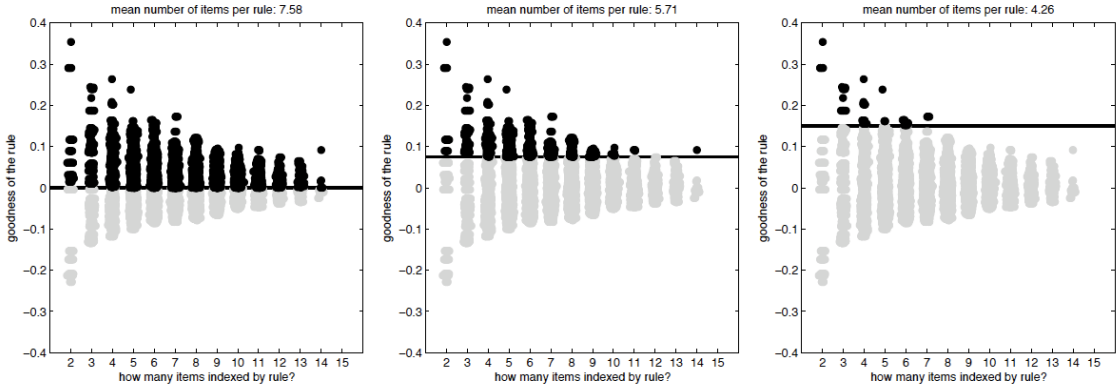


Figure 6. Illustration of the emergence of a sparsity bias. Each panel plots both the size m_h and goodness g_h of all 32768 possible categories for the kinship data. The solid line depicts the threshold γ; black dots correspond to rules that meet the threshold, grey dots are those that do not. The overall triangular shape of the data is a consequence of the central limit theorem, and so when the learner imposes the requirement that the admissable rules be good (i.e., $\gamma > 0$), a sparsity bias emerges.

Discussion

The analysis presented in this paper is necessarily somewhat simplified. We therefore briefly discuss a range of issues that it raises and the assumptions upon which it relies.

Structure in the world

Perhaps the most obvious simplification in this paper is that we make very minimal assumptions about the structure of the world: we assume that some stimuli are more similar than others, but there is no other source of structure built into the analysis. We also assume that the rules admitted to $\mathcal{H}$ are statistically independent of one another. Naturally, we expect that the real world is considerably more structured than these assumptions imply. This does not change our basic point, which is that only a very minimal amount of structure is actually *needed* to induce the sparsity bias and hence the PTS.

Nevertheless, it is interesting to consider in what ways the world might diverge from the assumptions we have made, and what this would mean for our analysis. Perhaps the most obvious point of divergence occurs when we relax the assumption that the rules are independent from one another. In that case, it might be possible for the learner to acquire structured background knowledge that allows him to bypass the capacity constraints and therefore make better queries based on this knowledge. When playing the everyday version of 20 questions, for instance, people typically start with "animal/mineral/vegetable" queries as a method for approximating globally optimal questions without explicitly representing the complete set of possible objects. In such cases it may be possible to improve on the positive test approach. However, the original Wason (1960) task – and many other typical situations, including scientific discovery – makes such a strategy difficult: the set of plausible rules is too large to hold in working memory, and there is no easy way to use global domain knowledge or exploit dependencies among the rules. Our analysis suggests that it is precisely in such cases that the positive test strategy approaches optimality.

Implicit falsification and the role of memory

One question worth considering is the difference between falsifying the explicit rules in $\mathcal{H}_E$ and those in the complementary set of *implicit* rules $\mathcal{H}_I$. This issue is reminiscent of early discussions of the rule-discovery task (Wetherick, 1962; Wason, 1962) in which a distinction was drawn between falsification of a hypothesis explicitly considered by a participant (falsification in $\mathcal{H}_E$), and the kind of implicit falsification that occurs because the oracle's responses are inconsistent with a great many rules, regardless of whether the learner has explicitly considered them (falsification in $\mathcal{H}_I$). The analysis in this paper treats both of these events as genuine falsifications, even though the learner would be unaware of implicit falsifications. The natural way in which this might happen is simply that these implicitly eliminated hypotheses are never considered, since they are inconsistent with previous observations. Nothing in the analysis requires learners to be aware of this, however; it merely requires that they not generate hypotheses that are inconsistent with the data so far.

Of course, this aspect of our analysis may be seen as a bit odd: why assume that the learner has limited memory capacity for hypotheses but perfect memory for the data? In many situations such an assumption is not unreasonable: for instance, it is trivially the case

that data are easily "remembered" by virtue of being written down, whereas the complete set of logically possible hypotheses is almost never easily accessible. In such cases there is a genuine asymmetry between data and hypotheses. However, since such arrangements are by no means universal, it is instructive to consider how the analysis would change if we relaxed the assumption that data are perfectly remembered. Our conjecture is that the same qualitative pattern would emerge, with a positive test strategy converging on the correct rule more quickly than other strategies. The main difference would be that all strategies would learn more slowly. To see why this would be the case, suppose that we (rather simplistically) formalize the notion of limited memory by assuming that some data are randomly deleted from memory. This would effectively be the same as if that data had never been seen: any hypotheses that would have been eliminated by such data would no longer be treated as having been falsified. However, since the analysis does not depend on any particular assumptions about what sort of data has or has not been seen, dropping data at random would not change the relative performance of different types of strategies. Only if the learner is more likely to forget data resulting from positive tests than negative tests would this analysis change; but if anything, the opposite assumption may be true (Taplin, 1975).

This does raise the broader question of what our approach presumes about the nature of memory. We assume that the learner constructs a space of plausible rules $\mathcal{H}$, from which a small number of explicit hypotheses $\mathcal{H}_E$ are selected and tested. One possible interpretation of such an assumption is that the larger hypothesis space $\mathcal{H}$ might correspond to concepts that are activated in long-term memory, while the explicit hypotheses $\mathcal{H}_E$ could correspond to the contents of working memory. However, we stress that this interpretation is by no means critical to the analysis: we adopted this "two-stage" construction primarily so that we could analyze hypothesis generation and hypothesis testing as two distinct problems. Introducing the idea that $\mathcal{H}$ corresponds to a set of "plausible" rules allowed us to show that $\mathcal{H}$ tends to be sparse in an analysis distinct from demonstrating that the PTS follows from this sparsity. Yet this construction is rather artificial; for instance, we could consider a situation in which long term memory has access to all 32768 rules in Figure 6 and generates explicit hypotheses $\mathcal{H}_E$ directly, but with a strong bias to prefer good rules, or in such a way as to place a high prior probability over the best rules. This would produce a similar result, although the full story in this case is slightly more complex and is beyond the scope of this paper.

Other extensions

One of the respects in which our analysis is less general than that of Klayman and Ha (1987) is that we have addressed only the "active learning" situation where the learner chooses the query item and the oracle provides answers to the query. The natural counterpart to this scenario – considered by Klayman and Ha (1987) but omitted here – is "observational learning", where the oracle generates either positive evidence (consistent with the true rule) or negative evidence (inconsistent with the true rule). The main reason we do not discuss this situation is that there exists a considerable body of work that bears on this problem already. The Bayesian theory of generalization (Tenenbaum & Griffiths, 2001; Sanjana & Tenenbaum, 2002; Navarro, Lee, Dry, & Schultz, 2008) is strongly premised on the notion that natural concepts tend to be sparse phenomena, for which positive evidence

is particularly powerful, and that people leverage off this to learn the meaning of words (Xu & Tenenbaum, 2007) among other things. Similarly, recent research on "pedagogical sampling" suggests that learning in both adults and children is sensitive to the nature of the choices made by the oracle (Shafto & Goodman, 2008; Bonawitz et al., 2009). And while early papers in computational learning theory seemed to suggest the children should not be able to learn languages from the positive-only evidence (Gold, 1967), more recent work (e.g., Muggleton, 1997; Chater & Vitànyi, 2007; Perfors, Tenenbaum, & Wonnacott, in press) suggests that learning from positive evidence can be quite powerful. In view of this existing literature and the fact that the classic decision-making results (Wason, 1960, 1968) relate primarily to the active learning scenario, we only discuss the active case in detail here.

Our analysis is also restricted to the rule-discovery problem. This problem is quite broad, covering a range of interesting problems in science, language learning and categorization, but there are other tasks that remain outside its scope. Most obviously, since our main goal was to consider the interaction between hypothesis generation and hypothesis testing, the link to the four-card selection task is complicated by the fact that it does not involve any hypothesis generation. This is not to say that there are no implications to be drawn, but in view of the fact that other analysis already cover the selection task in some detail (Oaksford & Chater, 1994) we do not address it here.

Capacity limited rationality

Since the analysis makes use of Bayesian theories of categorization (Anderson, 1991; Griffiths et al., 2008), hypothesis generation (Gettys & Fisher, 1979) and hypothesis testing (Klayman & Ha, 1987), it should be clear that we have focused primarily on the optimal solution to the computational problem facing the learner (Marr, 1982), and not on the processes by which such solutions may be reached. Even so, our analysis differs from a "standard" computational level analysis (e.g. Shepard, 1987; Tenenbaum & Griffiths, 2001) since we assume that processing constraints (probably in the form of working memory) that place a limit on the number of hypotheses one can explicitly consider play a central role in explaining human behavior. Thus, ours is a rational analysis in the sense that we show that positive testing is a *conditionally* optimal search strategy for a learner who has to deal with those constraints. It is similar in spirit to the "rational approximation" idea proposed by Sanborn, Griffiths, and Navarro (in press) and the "locally Bayesian learning" idea considered by Kruschke (2006). In fact, one way of characterizing the approach is as an attempt to incorporate architectural assumptions into rational analysis (see Anderson, 2007, for discussion). As we conceive it, the working memory constraint on the size of $\mathcal{H}_E$ and the long term memory constraint on the size of $\mathcal{H}$ are properties of the underlying cognitive architecture. Rather than ignore these ideas when modeling human learning, we instead treat the core ideas (e.g., working memory constraints) as fixed and then seek to solve the relevant computational problem (rule learning) conditional on this architecture. The result is a kind of rational analysis that still deals with universal computational problems (Shepard, 1987) but nevertheless is sensitive to the kinds of basic memory limitations (Miller, 1956; Cowan, 2001) that might apply to real-world learners.

Conclusion

We demonstrate two findings in this paper. In the realm of hypothesis testing, we show that when rules are sparse and the learner has access only to a small number of hypotheses, the PTS tends to be the best learning method (as in Klayman & Ha, 1987; Oaksford & Chater, 1994). In the realm of hypothesis generation, we suggest that sparse hypothesis spaces arise as a natural consequence of choosing sensible rules (this is implicit but not developed in some information-theoretic analyses, such as Cherubini et al. (2005) and Frank, Goodman, Lai, and Tenenbaum (2009)). While there is considerable existing evidence that people do represent the world sparsely, and the practical usefulness of sparse representations is evident in fields as diverse as machine learning, statistics, and neuroscience, we are not aware of any existing work that explicitly connects the two realms.

Taken together, the results provide a deeper theoretical explanation for the PTS. We envisage a learner who operates in a structured world and generates hypotheses in a Bayesian way (Gettys & Fisher, 1979), such that the prior over rules assigns higher probability to those rules that correspond to "good" family resemblance categories. Since good categories tend to be sparse, the learner tests them using the method that is optimal for sparse rules: the positive test strategy. Of course, this strategy is not perfect: no strategy can be when handling inductive problems, as the original Wason (1960) example illustrates. However, since the inevitability of inductive failures is considered to be a hard problem in philosophy (Hume, 1739/1898; Goodman, 1955, 1972), scientific discovery (Feyerabend, 1975) and computational learning (Wolpert & Macready, 1997), we might forgive human participants for following an optimal strategy that happens to fail in some cases.

Acknowledgements

Part of this work was presented at the 31st annual meeting of the Cognitive Science Society. DJN was supported by an Australian Research Fellowship (ARC grant DP0773794). We thank all of the people who commented on or reviewed the manuscript and the conference paper that it expands.

References

Adsit, D., & London, M. (1997). Effects of hypothesis generation on hypothesis testing in rule-discovery tasks. *The Journal of General Psychology*, *124*(1), 19–34.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Anderson, J. R. (2007). *How can the human mind exist in the physical universe?* Oxford: Oxford Univerity Press.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multi-dimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33-53.

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, *63*, 546-554.

Austerweil, J., & Griffiths, T. (2008). A rational analysis of confirmation with deterministic hypotheses. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Bonawitz, L., Shafto, P., Gweon, H., Chang, I., Katz, S., & Schulz, L. (2009). The double-edged sword of pedagogy: Modeling the effect of pedagogy on preschoolers' exploratory play. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Brown, A. S. (1974). Examination of hypothesis-sampling theory. *Psychological Bulletin*, *81*(11), 773 - 790.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Chater, N., & Vitànyi, P. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3), 135–163.

Cheng, P. W., & Holyoak, K. J. (1985). Causal reasoning schemas. *Cognitive Psychology*, *17*, 391-416.

Cherubini, P., Castelvecchio, E., & Cherubini, A. M. (2005). Generation of hypotheses in Wason's 2-4-6 task: An information theory approach. *The Quarterly Journal of Experimental Psychology*, *58A*(2), 309–332.

Chomsky, N. (1957). *Syntactic structures*. Berlin: Mouton.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *34*, 93–107.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-185.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., et al. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 1030-1048.

Doherty, M., Mynatt, C., Tweney, R., & Schiavo, M. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 111-121.

Dougherty, M., & Hunter, J. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, *113*(3), 263 – 282.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, *17*, 397–434.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, *38*, 467-474.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107-140.

Evans, J. S. B. T. (1972). Interpretation and 'matching bias' in a reasoning task. *Quarterly Journal of Experimental Psychology*, *24*, 193-199.

Evans, J. S. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning*, *4*, 45–82.

Evans, J. S. B. T., & Lynch, J. (1973). Matching bias in the selection task. *British Journal of Psychology*, *64*, 391–397.

Farris, H., & Revlin, R. (1989). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory and Cognition*, *17*(2), 221-232.

Feyerabend, P. (1975). *Against method*. London: New Left Books.

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Freedman, E. (1992). Scientific induction: Individual versus group processes and multiple hypotheses. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, 183–188.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 325-340.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance*, *24*, 93–110.

Gettys, C. F., Mehle, T., & Fisher, S. (1986). Plausibility assessments in hypothesis generation. *Organizational Behavior and Human Decision Processes*, *37*(1), 14 - 33.

Glushko, R. J. (1975). Pattern goodness and redundancy revisited: Multidimensional scaling and hierarchical cluster analysis. *Perception and Psychophysics*, *17*, 158-162.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447-474.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Goodman, N. (1972). *Problems and projects*. Bobbs-Merrill.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Goodnow, J. J., & Pettigrew, T. (1955). Effect of prior patterns of experience upon strategies and learning sets. *Journal of Experimental Psychology*, *49*, 381-389.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton, FL: Chapman and Hall.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for Bayesian Cognitive Science* (pp. 303–328). Oxford: Oxford University Press.

Huba, G. L., Wingard, J. A., & Bentler, P. M. (1981). A comparison of two latent variable causal models for adolescent drug use. *Journal of Personality and Social Psychology*, *40*(1), 180-193.

Hume, D. (1739/1898). *A treatise of human nature*. London: Ward Lock.

Johnson-Laird, P., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, *63*, 395-400.

Johnson-Laird, P., & Wason, P. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*(2), 134–148.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*, 59-99.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*, 10687–10692.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1–48.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211-228.

Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596–604.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 461 - 469.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3-36.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*, 677–699.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Levine, M. (1959). A model of hypothesis behavior in discrimination learning set. *Psychological review*, *66*, 353-.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Mahoney, M., & de Monbruen, B. (1977). Psychology of the scientist: An analysis of problem-solving bias. *Cognitive Therapy and Research*, *1*(3), 229–238.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

McDonald, J. (1990). Some situational determinants of hypothesis-testing strategies. *Journal of Experimental Social Psychology*, *26*, 255–274.

McKenzie, C. R. M. (2005). Judgment and decision making. In R. L. G. Koen Lamberts (Ed.), *Handbook of cognition* (pp. 321–338). Sage.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica*, *52*(1-2), 87 - 106.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Muggleton, S. (1997). Learning from positive data. In S. Muggleton (Ed.), *Inductive logic programming* (p. 358-376). Berlin: Springer.

Mynatt, C., Doherty, M., & Tweney, R. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The Quarterly Journal of Experimental Psychology*, *29*(1), 85–95.

Mynatt, C., Doherty, M., & Tweney, R. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, *30*, 395–406.

Nahinsky, I. D. (1970). A hypothesis sampling model for conjunctive concept identification. *Journal of Mathematical Psychology*, *7*, 293-316.

Navarro, D. J., & Lee, M. D. (2002). Commonalities and distinctions in featural stimulus representations. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 685–690). Mahwah, NJ: Lawrence Erlbaum.

Navarro, D. J., Lee, M. D., Dry, M. J., & Schultz, B. (2008). Extending and testing the Bayesian theory of generalization. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1746–1751). Austin, TX: Cognitive Science Society.

Nelson, J. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979 – 999.

Nelson, J., Tenenbaum, J., & Movellan, J. (2001). Active inference in concept learning. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nosofsky, R. M. (1984). Choice, similarity and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104-114.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53-79.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Perfors, A. F., & Navarro, D. J. (2009). Confirmation bias is rational when hypotheses are sparse. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Perfors, A. F., Tenenbaum, J., & Wonnacott, E. (in press). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*.

Platt, J. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Popper, K. R. (1935/1990). *The logic of scientific discovery*. Boston, MA: Unwin Hyman.

Rogers, T., & McClelland, J. (2004). *Semantic cognition*. Cambridge, MA: MIT Press.

Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–77). Hillsdale, NJ: Erlbaum.

Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, *10*, 489-502.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*, 94-101.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (in press). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*.

Sanjana, N., & Tenenbaum, J. B. (2002). Bayesian models of inductive generalization. *Advances in Neural Information Processing Systems*, *15*, 51–58.

Schervish, M. (1995). *Theory of statistics*. Springer-Verlag.
Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for pedagogical situations. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.
Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.
Shepard, R. N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87-123.
Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, *7*, 82-138.
Taplin, J. (1975). Evaluation of hypotheses in concept identification. *Memory and Cognition*, *3*, 85-96.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.
Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155 - 185.
Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.
Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*, 3-22.
Tweney, R., Doherty, M., Worner, M., Pliske, D., Mynatt, C., Gross, K., et al. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, *32*, 109-123.
Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.
Wason, P. (1962). Reply to wetherick. *The Quarterly Journal of Experimental Psychology*, *14*, 250.
Wason, P. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.
Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *The Quarterly Journal of Experimental Psychology*, *14*, 246-249.
Williams, G. F. (1971). A model of memory in concept learning. *Cognitive Psychology*, *2*, 158-184.
Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*, 67-82.
Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245.
Yama, H. (2001). Matching versus optimal data selection in the Wason selection task. *Thinking and Reasoning*, *7*, 295-311.

Appendix

Expected posterior entropy

Suppose the learner poses query x and obtains response r from the oracle, and now has posterior distribution over hypotheses $P(h|r,x)$. The entropy of this posterior distribution is given by

$$H(h|r,x) = -\sum_{h\in\mathcal{H}} P(h|r,x)\ln P(h|r,x) \quad (13)$$

$$= -\sum_{h\in\mathcal{H}^*} \frac{1}{n_c}\ln\frac{1}{n_c}. \quad (14)$$

where $\mathcal{H}^*$ contains only those hypotheses that are consistent with the oracle's response. Since there are n_c such hypotheses, the effect of the summation is simply to cancel the leftmost $1/n_c$ term, and hence

$$H(h|r,x) = -\ln \frac{1}{n_c} \quad (15)$$

$$= \ln n_c. \quad (16)$$

From the perspective of the learner who has not yet posed the query x, the response r is not known. The *expected* posterior entropy $E_r[H(h|r,x)]$ can therefore be found by averaging the entropy corresponding to a "yes" response and the entropy corresponding to a "no" response, each weighted by their expected probability (which is n_y/N for a "yes" response and n_n/N for a "no"). Thus, $E_r[H(h|r,x)]$ is given by:

$$E_r[H(h|r,x)] = E_r[\ln n_c] \quad (17)$$

$$= \frac{n_y}{N} \ln n_y + \frac{n_n}{N} \ln n_n. \quad (18)$$

This produces the result in Equation 3.

Expected posterior entropy for probabilistic rules

In the probabilistic case, some proportion of the items ϕ are exceptions to the rule, either because the oracle is fallible (or probabilistic), or because the situation involves learning a genuine rule plus exception structure. Assume that the learner is able to select query items x so as to control the expected probability that the oracle will answer "yes", and adopts a policy of keeping this probability at a constant value y (and note that $n_y = N \times y$). Then, consider the case where the learner has asked q queries, and received responses r. A hypothesis h that is consistent with c of these queries assigns probability to the oracle's responses as follows:

$$P(r|h,x) = \phi^{q-c}(1-\phi)^c \quad (19)$$

Since the distribution over hypotheses is initially uniform $P(h) \propto 1$, the posterior probability of this hypothesis after q queries have yielded responses r is

$$P(h|r,x) = \frac{\phi^{q-c}(1-\phi)^c}{\sum_{h' \in \mathcal{H}} \phi^{q-c'}(1-\phi)^{c'}} \quad (20)$$

$$= \frac{\phi^{q-c}(1-\phi)^c}{\sum_{c'=0}^{q} n_{c'} \phi^{q-c'}(1-\phi)^{c'}} \quad (21)$$

where c' denotes the number of responses correctly predicted by h', and $n_{c'}$ is the number of hypotheses that predicted exactly c' responses correctly. The exact value of n_c depends somewhat on exactly which rules are in $\mathcal{H}$ and what queries are made, but to a first approximation we can estimate what this number should be, by making use of the "yes probability" y and the sparsity θ. If independence holds, the probability α that a hypothesis agrees with the oracle's response is just:

$$\alpha = y\theta + (1-y)(1-\theta) \quad (22)$$

If so, then we would expect that the most likely value for n_c is

$$n_c = N\binom{q}{c}\alpha^c(1-\alpha)^{q-c} \tag{23}$$

To calculate the expected posterior entropy, we need to consider all possible oracular response r patterns over the q queries. However, the main quantity of interest is z, the number of "yes" responses that the oracle actually gave.

$$E[H] = \sum_{z=0}^{q} P(z|y,q)H(h|r) \tag{24}$$

where

$$P(z|y,q) = \binom{q}{z}y^z(1-y)^{q-z} \tag{25}$$

and the entropy is

$$H(h|r) = -\sum_{h\in\mathcal{H}} P(h|r,x)\ln P(h|r,x) \tag{26}$$

As before, we can simplify this expression by converting the sum over h to a sum over c and multiplying each term by n_c. Once this is done, we can substitute terms to compute the expected posterior entropies for the probabilistic case. It is this calculation that produced the plots in Figure 1b, where we set $q = 1$ so that the results can be compared to the deterministic case in Figure 1a, and convert the horizontal axis from y to n_y for the same reason.

On the sparsity of family resemblance categories

In this section we provide details for the calculation of expected sparsity as a function of threshold. Firstly, note that Equation 11 implies that if a category contains m_h items chosen at random, the probability that $g_h > \gamma$ is given by

$$P(g_h \geq \gamma|m_h) = \Phi\left(-\gamma\frac{2m_h(m_h-1)(M-m_h)}{2M-m_h-1}\right) \tag{27}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function for a standard unit normal variate. Applying Bayes' rule we obtain

$$P(m_h|g_h \geq \gamma, M) \propto P(g_h \geq \gamma|m_h)\Pr(m_h|M), \tag{28}$$

where $P(m_h|M)$ is just the proportion of the 2^m logically possible categories that include exactly m_h entities. This is calculated based on elementary combinatorics:

$$P(m_h|M) = \frac{M!}{m_h!(M-m_h)!}\frac{1}{2^M}. \tag{29}$$

This then allows us to calculate the expected sparsity θ_γ of the hypothesis space $\mathcal{H}$ as a function of the threshold γ:

$$\theta_\gamma = \frac{1}{M}E[m_h|g_h > \gamma, M] \tag{30}$$

$$= \frac{1}{M}\sum_{m_h=1}^{M} m_h P(m_h \,|\, g_h > \gamma, M). \tag{31}$$

It is this function that is plotted in the left panel of Figure 5.