Confirmation bias is rational when hypotheses are sparse

Amy F. Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide
Adelaide, SA 5005 Australia

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide
Adelaide, SA 5005 Australia

Abstract

We consider the common situation in which a reasoner must induce the rule that explains an observed sequence of data, but the hypothesis space of possible rules is not explicitly enumerated or identified; an example of this situation is the number game (Wason, 1960), or "twenty questions." We present mathematical optimality results showing that as long as hypotheses are *sparse* – that is, as long as rules, on average, tend to be true only for a small proportion of entities in the world – then confirmation bias is a near-optimal strategy. Experimental evidence suggests that at least in the domain of numbers, the sparsity assumption is reasonable.

Keywords: rational analysis; decision making; confirmation bias; information

Introduction

Humans are constantly confronted with situations in which they must induce the underlying process or rule that generated the data they see. Children learning language must infer an underlying grammar on the basis of the sentences they hear; scientists must infer theories on the basis of the data they observe; and people trying to understand the world must infer explanations on the basis of the experience they have. One might expect that rational learners would approach this situation by evaluating hypotheses on the basis of both confirming and disconfirming evidence. However, one of the most well-studied and well-supported results in decision science demonstrates that, in a variety of situations, people tend to only seek after confirming evidence; this is known as the *confirmation bias* (see Nickerson (1998) for an overview).

The confirmation bias can cover a variety of situations. It includes times in which people are motivated to support or believe in a pet theory, perhaps for emotional reasons, and thus discount evidence that would falsify it (e.g., Matlin & Stang, 1978). It also includes times in which people overweight confirmatory evidence (e.g., Gilovich, 1983) or let their prior biases affect the evidence they see (e.g., Kuhn, 1989). We focus here on the hypothesis selection aspect of the confirmation bias: the tendency for people who are trying to determine which of a number of hypotheses is correct to ask questions that will get a "yes" response if the hypothesis currently under consideration is true (e.g., Mynatt, Doherty, & Tweney, 1978; Wason, 1960, 1968).

One classic example of this occurs in a task known as the Wason Selection task, in which participants are shown four cards with letters on one side and numbers on the other (Wason, 1968). They are asked to evaluate the truth of a rule of the form IF P, THEN Q where p = "there is a vowel on one side of the card" and q = "there is an even number on the other." The four cards shown the participants consist of one showing a vowel (a p card), one showing a consonant (a $\neg p$ card), one showing an even number (a q card), and one showing an odd number (a $\neg q$ card). Asked which card(s) they would flip in order to evaluate the truth of the rule, participants prefer to flip the p and q cards, even though logically they should select p and $\neg q$ (Johnson-Laird & Wason, 1970). This is a kind of confirmation bias because people seek out evidence that would confirm components of the rule (p and q), even though to evaluate the rule as a whole, both confirmatory and disconfirmatory evidence are relevant.

Although this tendency can be ameliorated under some conditions (e.g., Cosmides, 1989), it is rarely completely eliminated. However, some (Oaksford & Chater, 1994) have argued that the confirmation strategy in the Wason Selection Task can be seen as rational, if one assumes that the goal of hypothesis testers is to perform queries (i.e., select cards) that maximize the expected information gain in deciding between two hypotheses. The two hypotheses in the Wason Selection Task are (a) IF P THEN Q and (b) P AND Q ARE INDEPENDENT. Oaksford and Chater (1994) show that as long as people assume that the properties described in p and q are rare, then the query that maximizes information gain is the p card, and the next best query is the q card – and this is precisely what people do.

This is a provocative result, but it only applies to a subset of the interesting cases confronted by people. It is quite common (as in, for instance, most of the examples that we opened with) for people to have to choose between more than two hypotheses, many of which may not be explicitly identified or enumerated. For children learning language, scientists forming theories, and people forming causal explanations of the world – as well as many other situations – the space of potential hypotheses is infinite or near-infinite, and the hypotheses are not constrained to be of the form IF p, THEN q. This creates a much more difficult problem – one where the analysis of Oaksford and Chater (1994) may not apply, since their analysis requires that the hypotheses be enumerated and identified so that they might be compared using Bayes' Rule.

Yet people still show a confirmation bias in situations with potentially infinite hypothesis spaces with unenumerated hypotheses. In one task, also initially invented by Wason (1960), participants were asked to try to guess the rule that defines a sequence of three numbers. The participants suggested triads and received feedback about whether those triads were acceptable under the rule, which was NUMBERS ARE INCREASING (such that 2-4-6 would be acceptable but 4-2-6 would not be). They often failed to infer the rule, whilst still finding one that was consistent with all of the data they had seen (e.g., INCREASING POWERS OF TWO). Notably, participants who identified this sort of "subset rule" failed to try to disconfirm it by suggesting triads that would be unacceptable if it were the case (such as 2-4-6); instead they suggested triads that were predicted by the rule (such as 2-4-8).

This task is much more analogous to the sort of situations that face human learners and scientists all of the time: situations with many possible hypotheses (if not an infinite number), where not all are explicitly enumerated or considered at once. In this paper we evaluate this situation, and show that given certain assumptions about the nature of the hypotheses, a confirmation strategy is rational: it will allow the learner to confirm or disconfirm the current hypothesis with the least number of queries by maximising the average information gain of each query. We further provide reason to believe that the required assumption – that hypotheses are "sparse" – is reasonable in at least some situations.

The rational basis of confirmation bias

The rule learning problem that Wason (1960) discussed is a variant of the popular *twenty questions game*. In the everyday life version of the game, one player (who we will call the "oracle") thinks of an object, and the other player (the "learner") can pose *queries* to the oracle. These queries must take the form of a "yes or no" question, and the learner's goal is to ask questions that allow the object to be identified as quickly as possible. The formal specification of this task is:

> Suppose we have a collection of n entities in some domain $\mathcal{X}$, and a hypothesis space $\mathcal{R} = (r_1, \ldots, r_m)$ consisting of m possible rules, each of which is equally plausible *a priori*. Each such rule $r_i(\cdot)$ is a function that picks out a set of n_i rule-consistent entities for which $r_i(x) = 1$, leaving a set of $n - n_i$ rule-inconsistent entities for which $r_i(x) = 0$. The learner encounters an oracle O to which queries $x \in \mathcal{X}$ may be put; and yields answers $O(x)$ based on some unknown rule $r \in \mathcal{R}$. The learner's goal is to generate query items x in such a way that the nature of the oracle function $O(\cdot)$ can be inferred with as few queries as possible.

Bayesian learning and optimal queries

Since each of the m rules is considered equally likely *a priori*, the prior probability that the oracle function corresponds to the ith rule (denoted $O \to r_i$) is given by $\Pr(O \to r_i) = 1/m$. For any query item $x \in \mathcal{X}$ chosen by the learner and put to the oracle, we obtain the response $O(x)$. If we treat the rule r_i as a possible model for the oracle, then we can write down a simple likelihood function that assigns

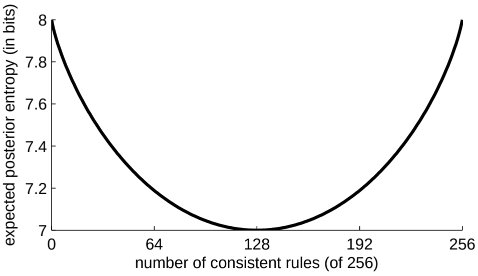

Figure 1: Expected posterior entropy of the oracle as a function of the query item. In this figure, we assume that there are $m = 256$ possible rules, meaning that the learner needs to uncover 8 bits of information to identify the rule. If a query is true of none or all of the rules, the learner gets no information from the reply and there remain 8 bits to find. Only if the query item is true for exactly half of the rules will the oracle's answer be maximally informative, and thus convey a complete bit of information.

$$\Pr(O(x) \mid O \to r_i) = \begin{cases} 1 & \text{if } O(x) = r_i(x) \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

Via Bayes' theorem, the posterior probability that the oracle function corresponds to the ith rule r_i is given by

$$\Pr(O \to r_i \mid O(x)) = \frac{\Pr(O(x) \mid O \to r_i) \Pr(O \to r_i)}{\sum_j \Pr(O(x) \mid O \to r_j) \Pr(O \to r_j)}. \tag{2}$$

Because $\Pr(O \to r_i)$ is a constant and $\Pr(O(x) \mid O \to r_i)$ is a binary function, the effect of receiving the information $O(x)$ is simple: all rules that are inconsistent with the oracle's answer are eliminated, and all remaining rules are still equally plausible. Formally, the degree of belief that the learner should have in rule r_i is:

$$\begin{array}{c|cc} & r_i(x) = 1 & r_i(x) = 0 \\ \hline O(x) = 1 & \dfrac{1}{m(x)} & 0 \\ O(x) = 0 & 0 & \dfrac{1}{m(\neg x)} \end{array} , \tag{3}$$

where $m(x)$ counts the number of rules $r \in \mathcal{R}$ that produces an affirmative response to $r(x) = 1$ the query, and $m(\neg x)$ counts the rules that return a negative response $r(x) = 0$. As more queries are made, the number of rules that are consistent with all of the oracle's answers will diminish, and eventually the leaner can correctly identify the rule.

How should the learner choose the next query item x? If the aim is to identify the rule as quickly as possible, a rational learner should choose the item x that minimizes the *expected posterior entropy* of their beliefs about the identity of the oracle. That is, she should pick the x that is expected to return the most information about the true rule (see MacKay, 2003). A formal derivation of the expected entropy is given in Equations 4–7 (next page), but the important thing to note is that the entropy is minimized when $m(x) = m(\neg x) = m/2$. This is

$$\begin{aligned} E[H(O|x)] &= -E_{O(x)}\left[\sum_i \Pr(O \rightarrow r_i \mid O(x)) \ln \Pr(O \rightarrow r_i \mid O(x))\right] && (4) \\ &= -\Pr(O(x)=0)\left(\sum_i \Pr(O \rightarrow r_i \mid O(x)=0) \ln \Pr(O \rightarrow r_i \mid O(x)=0)\right) && \\ &\quad -\Pr(O(x)=1)\left(\sum_i \Pr(O \rightarrow r_i \mid O(x)=1) \ln \Pr(O \rightarrow r_i \mid O(x)=1)\right) && (5) \\ &= -\frac{m(\neg x)}{m}\left(\sum_{i|r_i(x)=0} \frac{1}{m(\neg x)} \ln \frac{1}{m(\neg x)}\right) - \frac{m(x)}{m}\left(\sum_{i|r_i(x)=1} \frac{1}{m(x)} \ln \frac{1}{m(x)}\right) && (6) \\ &= \frac{m(\neg x)\ln m(\neg x) + m(x)\ln m(x)}{m} && (7) \end{aligned}$$

illustrated in Figure 1: the optimal query x is one that is true for exactly half of the not-yet-eliminated hypotheses. This is the "bisection" search method that people intuitively prefer to use in simpler situations; if asked to identify a number in the range 0-100 using only "greater than/less than" questions, most people tend to use 50 as the first query.

Optimal queries given partial knowledge

From the discussion in the previous section, if we have a *known* collection of rules $\mathcal{R}$ and one of them corresponds to the oracle O, then the bisection approach is the optimal search method. Under these circumstances, there ought to be no confirmation bias – the ideal query should falsify exactly as many hypotheses as it confirms. However, this situation does not really match Wason's (1960) task or the twenty questions game generally. In most interesting cases, the number of possible rules m is very large, and so the complete rule set $\mathcal{R}$ will *not* be known to the learner when he or she is attempting to select a query item x. Instead, only some (usually small) subset of the possible rules $\mathcal{R}_E$ are likely to be explicit and available, with the remainder $\mathcal{R}_I$ remaining implicit. Under those circumstances, what should a rational learner do?

To help answer this question, we consider the slightly simplified case where the rules are independent of one another (so that in general, knowing rule r_i does not tell you anything about rule r_j). We also make the critical assumption that the rules are *sparse*. The sparsity assumption states that most (though not necessarily all) potential rules are only true for a small proportion of entities in the world. For instance, when considering possible rules that might be satisfied by the numbers between 1 and 1000, the rule IS DIVISIBLE BY 12 is sparse, whereas IS EVEN and IS NOT DIVISIBLE BY 12 are not sparse. Because every sparse rule has a non-sparse complement, sparsity does not logically hold, but we provide evidence in the next section that the rules that people are likely to be interested in will tend to be sparse.

Under these circumstance, the learner can assume that the prior probability of any rule returning "yes" will be given by the average sparsity of all of the hypotheses θ. As before, if all rules are known (i.e., $\mathcal{R}_E = \mathcal{R}$) then the bisection method is optimal, and is presumably an achievable strategy. At the other extreme, if no rules are known (i.e., $\mathcal{R}_E = \emptyset$), then the learner has no control over the effectiveness of the query item x. From his or her perspective, the number of rules $m(x)$ that can be falsified by using the query x is on average $m\theta$, with the actual number being binomially distributed, as follows:

$$m(x) \sim \text{Binomial}(\theta, m). \qquad (8)$$

However, because the total number of rules m tends to be large and the rules themselves tend to be sparse ($\theta \ll \frac{1}{2}$), the query will almost certainly be suboptimal, since $m(x)$ will almost certainly be much smaller than $m/2$. In short, in a sparse world, random queries will falsify more hypotheses than they confirm, and will hence lead to suboptimal learning.

With this in mind, we now consider the situation in which the learner has partial knowledge: some rules are known and others are not. The learner's goal is still to identify the oracle $O \in \mathcal{R}$ as quickly as possible, but only some subset of rules $\mathcal{R}_E$ can be used to guide the choice of x.[1] Even if the implicit rules $\mathcal{R}_I$ are not enumerated, the oracle's response $O(x)$ will still be informative about them: for instance, even if the learner is not explicitly thinking of the possibility that the rule is IS A FIBONACCI NUMBER when they select $x = 12$ as a query, an oracle response of "yes" still counts as a falsification of that hypothesis (though it may take the learner some time to notice this). If there are m_E explicit rules and m_I implicit rules, the independence assumption implies that any query item x chosen using the information in $\mathcal{R}_E$ is still a random binomial draw with respect to $\mathcal{R}_I$, with

$$m_I(x) \sim \text{Binomial}(\theta, m_I). \qquad (9)$$

Thus, by the sparsity assumption, any query x chosen by a consideration of $\mathcal{R}_E$ will (with high probability) be inefficient with respect to $\mathcal{R}_I$, since $m_I(x)$ is likely to be much

[1] We also treat the value of θ as known in order to simplify the discussion, but these results would hold more generally even if it is not known, as long as θ is actually sparse.

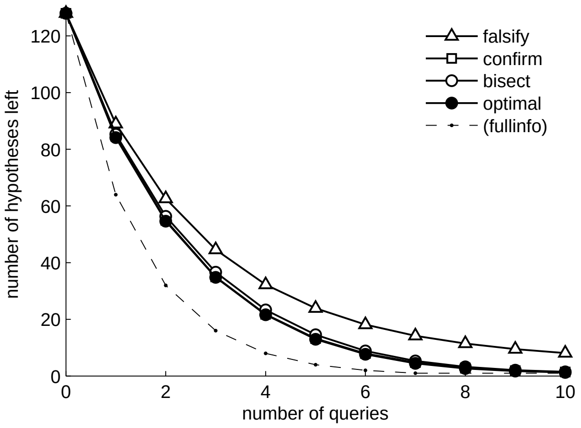

Figure 2: Rates of hypothesis elimination in a twenty-questions game involving 128 (random) hypotheses, 50 possible queries and sparsity of 0.2. In this case, it is assumed that the learner can consider seven hypothesis at a time ($m_E = 7$), and keeps each hypothesis until it is falsified. We consider four strategies: confirmation, falsification, limited bisection of the 7 explicit hypotheses, and the optimal strategy (which favors confirmation at the beginning and bisection as the number of explicit hypotheses approaches the total number of hypotheses remaining). The confirmation strategy eliminates hypotheses faster than the falsification strategy. The dashed line shows the (unachievable) rate of elimination when a perfect bisection method (over a fully enumerated hypothesis space) is followed.

less than $m_I/2$, and hence biased in the direction of falsification. If the number of implicit, unrepresented rules is much larger than the set of represented rules, the learner is almost certainly better off trying to pursue a strategy that is highly confirmation-biased with respect to $\mathcal{R}_E$ in order to counteract the falsification bias among the hidden hypotheses.

In short: if the world is sparse and the learner does not have access to all relevant hypotheses at all times, it is optimal to have a confirmation bias with respect to the set of hypotheses that the learner does have access to. However, as the number of implicit hypotheses decreases the extent of the bias should reduce, and when all hypotheses are explicit, the bisection strategy becomes optimal.

Is the sparsity assumption reasonable?

On a purely descriptive level, the sparsity assumption does not match the actual distribution of possible number rules. There are an equivalent number of sparse and non-sparse hypotheses, since any sparse hypothesis is matched by a complementary non-sparse hypothesis: the hypothesis PRIME NUMBERS is matched by the hypothesis NON-PRIME NUMBERS, and so on. In that sense, then, the sparsity assumption does not appear justified.

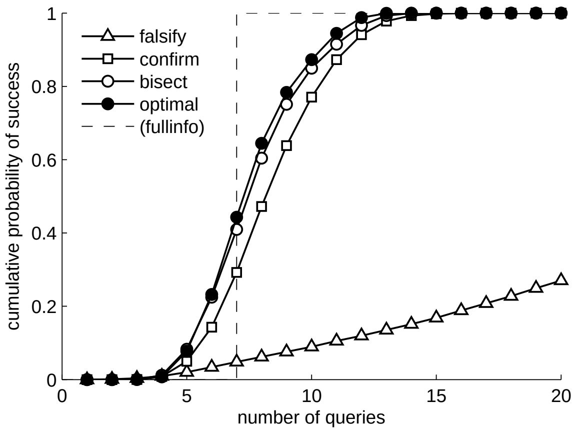

Figure 3: A different perspective on the same scenario as in Figure 2. Each plot shows the probability that the learner can identify the rule within some number of queries. Interestingly, since the full-information bisection method always solves the problem in exactly seven guesses, a learner who is restricted to fewer than this number would be better off trying to confirm. Although the limited bisection strategy slightly outperforms the confirmation strategy, both far outperform falsification.

However, the hypotheses people tend to entertain, or weight as *a priori* more probable, may not map directly onto the set of all possible allowable hypotheses. In the numbers game, it seems reasonable to consider the hypotheses that are easily expressible in terms of the mental representations that people use to encode numbers. While some natural features of numbers yield non-sparse hypotheses (e.g., evenness or oddness), intuitively it seems clear that most are sparse. Sometimes this may be for arbitrary reasons related to notation and other factors: base 10 orthography means that rules like CONTAINS A PARTICULAR NUMERAL only indexes a fraction of the numbers in the domain. Other times, this may be for deeper reasons: MULTIPLES OF *n* cannot be dense, nor can rules about NUMBERS RAISED TO THE POWER *n*.

One might object that each of these sparse rules r_S has a non-sparse analogue, easily represented mentally as $\neg r_S$, or that less-sparse rules could easily be encoded as disjunctions of sparser rules, e.g. $r_{S1} \vee r_{S2}$. However, as long as one assumes that these rules have longer encoding lengths in whatever mental representational language people use for numbers (i.e., the mind still has to encode the logical operators $\neg$, $\vee$, $\wedge$ etc), and that prior probability is higher for shorter encodings, then these sorts of rules should be given less weight by the learner (Solomonoff, 1964).[2] Intuitively, then, it is plausible to assume that (at least in the number domain) the hypotheses that people most likely represent and use tend to be sparse, that these sparse hypotheses would have shorter codelengths, and therefore would be considered *a priori* more probable than non-sparse hypotheses.

In the next section we empirically evaluate whether this intuition is an accurate one.

[2] To be fair, the same observation would apply to conjunctive rules of the form $r_{S1} \wedge r_{S2}$, which would be even sparser than the original r_{S1} and r_{S2} hypotheses; but this would simply counterbalance the disjunctive rules.

Experiment

We presented 16 participants with a paper-and-pencil task in which they were asked to list all of the possible rules they could think of that could apply to numbers on the domain of [1,1000]. We do not assume that the rules presented perfectly reflect all of the rules that people implicitly or explicitly consider; indeed, one of the assumptions of our analysis has been that many rules are unenumerated and unidentified at the time of making the query. Nevertheless, the experiment was designed to provide at least some empirical data about what sort of rules people tend to consider most.

In order to ensure that we did not bias subjects to prefer any kind of rule over any other, the task was made clear using as an example rules over the domain of the alphabet [A, Z]. These rules were described to subjects in the following way:

> If you were asked to guess a rule that could pick out a set of letters from the alphabet, many different rules might occur to you. If you were asked to list the rules you thought of, and rank which ones you think are most or least likely, you might come up with the items [we consider] below. While everyone might come up with a slightly different list, this illustrates the kind of thing we are talking about. Some rules, like ALL VOWELS or ALL CONSONANTS, are fairly obvious. Other rules are entire classes of rules, like RHYMES WITH <SOME WORD>: this includes the rule RHYMES WITH "BEE" (which includes B, C, D, and others, but not F) as well as the rule RHYMES WITH "BAY" (which includes A, J, and K) as well as many others. And other rules might be very strange and unlikely, but still possible, like the rule ALL LETTERS THAT ARE PRONOUNCED BY CLOSING THE LIPS (B, P, or M).

We then told subjects that we were interested in how people think about the rules that pick out sets of numbers, and asked them to list all of the rules they could think of that "pick out some set of numbers from the ones between 1 and 1000."

Results

The 16 participants produced a total of 70 distinct rules (maximum 19, minimum 3). There was substantial overlap on several rules, but most participants produced rules that none others suggested. Figure 4 displays all of the rules suggested by at least two distinct people.

The degree of agreement between subjects about whether to list a rule appears to give a measure of how likely each rule may be, *a priori*. While this is not conclusive, it does seem reasonable to think that the rules EVEN NUMBERS and ODD NUMBERS are weighted more highly in people's mental representations than rules like PERFECT CUBES. In addition, several of the listed rules are technically, proper subsets of each other: EVEN/ODD as well as MULTIPLES OF 5 (or 10) are technically all examples of the rule MULTIPLES OF X. However, most participants listed these as separate rules, and many noted specifically that although they recognized that some were subsets, they listed them separately because it "felt like they were different." Since we are interested in characterising how people actually represent number rules, we treat them as distinct as well.

Do these results lend empirical support to the assumption that hypotheses in this domain are sparse? To determine this, we calculated the estimated sparsity θ for each of the rules listed by more than one participant; these sparsity values are shown in Figure 5. For most rules, this calculation was straightforward: 50% of the numbers on the domain [1,1000] are even, so the sparsity of that rule was 0.5; and there are 168 prime numbers below 1000, so the sparsity of PRIME NUMBERS was 0.168. Three of the rules were impossible to calculate the sparseness of, since they make reference to idiosyncratic properties of the participants (e.g., ADDRESSES); these are indicated visually with a bar depicting a negative sparsity value. It is probable that all three are fairly sparse, since most people do not know hundreds of distinct addresses, dates, or phone numbers.

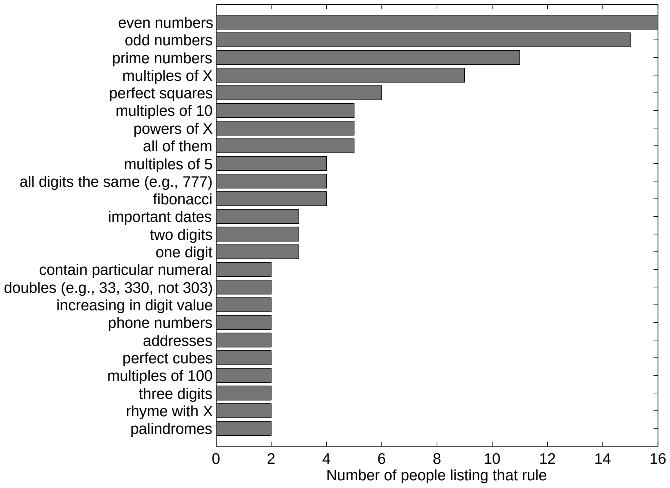

Figure 4: Each of the rules listed by at least two participants. The bar graph shows the number of participants suggesting each rule, out of 16 participants total.

A few of the rules required less straightforward calculation, because they implicitly assumed a distribution over entire classes of rules: the rules MULTIPLES OF X and POWERS OF X are two examples. For each, we calculated sparsity by assuming that it was a weighted measure reflecting the ten smallest X. For instance, for the rule MULTIPLES OF X, we considered the 10 sub-rules MULTIPLES OF 3, MULTIPLES OF 4, . . ., MULTIPLES OF 14.[3] We weighted each inversely proportional to its normalized rank, so that MULTIPLES OF 3 was weighted most highly and MULTIPLES OF 14 was weighted least. The average of these is the sparsity. This results in a sparsity that is higher than it would be if all of the sub-rules were weighted equivalently.

It is evident that the clear majority of the rules (83%) have sparsity values of 0.2 or less. Calculating average sparsity yields a $\theta = 0.2087$, and weighting each rule inversely proportional to the number of people listing it yields $\theta = 0.0989$. Either way, this provides some degree of empirical support for the idea that people's hypotheses in the "number game"

[3] This list does have exactly 10 rules: we excluded MULTIPLES OF 5 and MULTIPLES OF 10 from the list, since they were listed separately elsewhere.

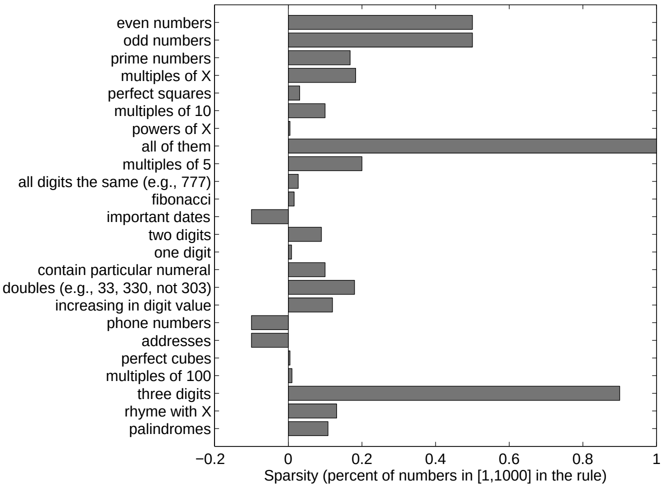

Figure 5: Estimated sparsity values for each of the rules listed by at least two participants. Sparsity for three rules could not be calculated; this is indicated visually with a negative value. Most of the other rules have sparsity values of less than 0.2, supporting the assumption that hypotheses in this domain tend to be sparse.

are, indeed, quite sparse.

This result still appears to be the case even if we include idiosyncratic rules (the ones listed by only one participant each). Although their idiosyncraticness makes some of them more difficult to analyse, an inspection of them suggests that most of them are sparse as well. Representative examples include rules like CURRENCY DENOMINATIONS, NUMBERS WRITTEN WITH STRAIGHT LINES ONLY, TOP 10 NBA SCORING AVERAGES, NUMBERS THAT SPELL WORDS ON THE TELEPHONE DIAL, LIVELY-FEELING NUMBERS, and PERFECT NUMBERS. It seems reasonable to think that the sparsity of these hypotheses does not differ qualitatively from the sparsity of the rules that there is more agreement on.

Discussion

Although we only provide empirical evidence about the sparsity of hypotheses in the domain of number rules, it may be reasonable in other domains as well. Scientific hypotheses are especially valued for their sparsity: a theory is valuable to the extent that it makes specific predictions, and one that licenses nearly any behavior is not very useful. Much work remains to be done to determine whether sparsity is generally applicable, and whether people show a reduced confirmation bias in domains where it has been shown not to be.

More generally, in light of the strongly held view that confirmation bias is a robust indicator of human irrationality, the optimality results in the paper may come as a surprise. If the confirmation strategy works so well, why was this not evident when Wason (1960) first documented the phenomenon? Part of the reason, we suspect, is that there are at least three important cases in which a confirmation bias is a poor strategy even when the sparsity assumption is met. The most trivial include instances like those we discussed in the introduction, when "confirmation bias" means overweighting confirming evidence, interpreting new data in light of one's prior beliefs, or refusing to engage with disconfirming evidence for emotional reasons. A second case is one we already noted, when the set of hypotheses is fully enumerated, and a bisection strategy is optimal.

The third case may occur when the assumption that rules are independent is no longer true: in many cases it may be possible to exploit correlations between hypotheses to speed the search process. For instance, when playing the everyday version of twenty questions, people typically start with animal/mineral/vegetable queries. Even though people may not be able to explicitly represent all the possible answers, they know that the structure of the domain is such that these questions provide a good approximation to the globally-optimal bisection strategy. In such cases, it is possible to improve on the confirmation approach. However, the original Wason (1960) task makes this strategy difficult – the set of "plausible rules about sequences of numbers" is too large to hold in working memory, and does not naturally allow any easy way to use global domain knowledge or exploit any dependencies among the sparse rules. It is in exactly such cases that confirmation bias is optimal, because there are so many ways to be wrong and so few ways to be right. In such instances, the learner will discover that the world has a falsification bias, and a confirmation strategy presents the best way to fix it.

Acknowledgments

AFP and DJN contributed equally to this work. We thank Xin Wei Sim and Steven Langford for their assistance with data collection. DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794).

References

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.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Psychology and Social Psychology*, *44*(6), 1110–1126.

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

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*(4).

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

Matlin, M., & Stang, D. (1978). *The Pollyanna principle: Selectivity in language, memory, and thought*. Cambridge, MA: Shenkman.

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.

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

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

Solomonoff, R. (1964). A formal theory of inductive inference, parts 1 and 2. *Information and Control*, *7*(1–22), 224-254.

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. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.