Journal of Mathematical Psychology 77 (2017) 10–20

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Online learning of symbolic concepts

Pratiksha Thaker [a], Joshua B. Tenenbaum [b], Samuel J. Gershman [c,*]

[a] Department of Computer Science, Stanford University, United States
[b] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States
[c] Department of Psychology and Center for Brain Science, Harvard University, United States

HIGHLIGHTS

- A novel variant of the number game is studied.
- An online approximate Bayesian model captures order effects in concept learning.
- Placing people under cognitive load strengthens the order effect.

ARTICLE INFO

Article history:
Received 30 July 2016
Received in revised form
28 November 2016

Keywords:
Bayesian models
Concept learning
Bounded rationality
Particle filtering

ABSTRACT

Learning complex symbolic concepts requires a rich hypothesis space, but exploring such spaces is intractable. We describe how sampling algorithms can be brought to bear on this problem, leading to the prediction that humans will exhibit the same failure modes as sampling algorithms. In particular, we show that humans get stuck in "garden paths"—initially promising hypotheses that turn out to be sub-optimal in light of subsequent data. Susceptibility to garden paths is sensitive to the availability of cognitive resources. These phenomena are well-explained by a Bayesian model in which humans stochastically update a sample-based representation of the posterior over a compositional hypothesis space. Our model provides a framework for understanding "bounded rationality" in symbolic concept learning.

1. Introduction

One of the most remarkable characteristics of human cognition is the ability to learn symbolic concepts from very sparse data. For example, after being shown the numbers {60, 80, 10, 30} drawn from a set of numbers between 1 and 100, humans will confidently infer the set to be "multiples of ten" (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001). This kind of strong generalization requires a hypothesis space rich enough to express a wide variety of concepts, as well as a mechanism for efficiently exploring the hypothesis space and evaluating candidate concepts. A conundrum at the heart of concept learning is that these two requirements are at odds with one another: The richer the hypothesis space, the harder it is to efficiently explore. This is especially true for compositional hypothesis spaces (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kemp, 2012; Piantadosi, Tenenbaum, & Goodman, 2010), where the number of possible concepts is exponential in the number of primitives. Moreover, Bayesian approaches to concept learning assert that humans represent a probability distribution over the entire hypothesis space (Shepard, 1987; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001). These considerations bring the issue of computational tractability to the foreground.

Previous treatments of symbolic concept learning have primarily focused either on abstract rational analysis without detailed mechanistic commitments (Feldman, 2000; Kemp, 2012; Piantadosi et al., 2010; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001) or on mechanistic models without a clear connection to rational inductive principles (Goodwin & Johnson-Laird, 2011; Kruschke et al., 1992; Nosofsky, Palmeri, & McKinley, 1994). Goodman et al. (2008) used a compositional grammar to model boolean concept learning, and presented provisional evidence that participants adhere to rational inductive principles only approximately: Behavior was best explained by assuming that humans make their responses using one or a few samples from the posterior distribution over concepts. Hypothesis sampling has become an important bridge between rational analyses and process models (see Griffiths, Vul, & Sanborn, 2012, for a review), with applications to vision (Gershman, Vul, & Tenenbaum, 2012; Moreno-Bote, Knill, & Pouget,

* Correspondence to: Department of Psychology, Harvard University, 52 Oxford St., room 295.05, Cambridge, MA 02138, United States.
E-mail address: gershman@fas.harvard.edu (S.J. Gershman).

http://dx.doi.org/10.1016/j.jmp.2017.01.002
0022-2496/

2011; Vul, Frank, Alvarez, & Tenenbaum, 2009; Wozny, Beierholm, & Shams, 2010), theory learning (Denison, Bonawitz, Gopnik, & Griffiths, 2013; Ullman, Goodman, & Tenenbaum, 2012) and categorization (Sanborn, Griffiths, & Navarro, 2010), among others. Empirical evidence for hypothesis sampling will be reviewed in the General Discussion.

Hypothesis sampling provides a simple model of cognitive limitations (in terms of how many samples are used) while instantiating a theoretically sound mechanism for approximating Bayesian inference (Vul, Goodman, Griffiths, & Tenenbaum, 2014). In particular, many hypothesis sampling models can be viewed as Monte Carlo methods, which are widely used in statistics and machine learning due to their flexibility and theoretical properties (Robert & Casella, 2004). Previous work on concept learning in compositional hypothesis spaces used Markov chain Monte Carlo (MCMC) algorithms to generate samples (Goodman et al., 2008; Piantadosi et al., 2010); since these algorithms evaluate hypotheses over the entire data set at each iteration, they are cognitively implausible for tasks in which data are presented sequentially and presumably processed online (as in many domains, like word learning or multiple-object tracking).

This paper investigates a cognitively plausible sampling algorithm for performing online inference over a compositional hypothesis space of number concepts. Our starting point is the "number game" described in Tenenbaum (1999). In this experiment, participants were presented with a set of integers generated from a number concept (a subset of numbers between 1 and 100), such as "all powers of 2" or "all numbers between 40 and 60". Participants were then asked to judge, for several other numbers, the probability that each was generated from the same subset as the examples presented. Tenenbaum (1999) showed that generalization patterns in this experiment were consistent with a Bayesian model of concept learning (described in more detail below). While the space of number concepts is very large, Tenenbaum's model constrained the hypothesis space to a small number of intuitively plausible concepts.

We will consider a richer space of compositional concepts, and postulate a form of hypothesis sampling as a theory of how humans explore this hypothesis space. In particular, we argue that humans use an online hypothesis sampling algorithm called *particle filtering* that entertains multiple hypotheses ("particles") and continually reweights the particles as new data are observed. This algorithm has previously been used to explain aspects of multiple object tracking (Vul et al., 2009), category learning (Sanborn et al., 2010), change detection (Brown & Steyvers, 2009), word segmentation (Frank, Goldwater, Griffiths, & Tenenbaum, 2010), and reinforcement learning (Daw & Courville, 2007; Yi, Steyvers, & Lee, 2009). While most of this previous work has focused on hypothesis spaces with relatively simple representational structure (e.g., mixture models), our goal is to provide empirical constraints on hypothesis sampling in more complex symbolic spaces.

One implication of hypothesis sampling is that when faced with complex or ambiguous example sets in the number game, participants might fail to infer some concepts that have high posterior probability. We speculated that this might happen if examples are presented to participants sequentially, such that the early examples favor one concept, but the later examples tilt the posterior in favor of a different concept. If conditionally unlikely samples are eliminated during hypothesis sampling (an operation known as "resampling"), the early, sub-optimal concept will prevail. This is analogous to "garden path" sentences in psycholinguistics (e.g., "we painted the walls with cracks") which are difficult for humans to parse (MacDonald, 1994). Levy, Reali, and Griffiths (2009) modeled garden path effects with hypothesis sampling by assuming that the correct parse was eliminated from the sample set early on during sentence processing. We adapted this model to number concept learning, and constructed example sequences which would lead the model to show garden path effects. We then conducted experiments with humans to test whether participants show the same effects.

The plan of the paper is as follows. Section 2 summarizes the Bayesian framework for concept learning developed by Tenenbaum (1999), and introduces a hypothesis sampling algorithm for approximate inference. Section 3 reports an experiment with human participants playing a sequential concept learning game. We show how the hypothesis sampling algorithm provides a rational process account of order effects and cognitive load manipulations in the game. Section 4 concludes the paper with a discussion of related work and future directions.

2. A Bayesian framework for concept learning

In this section, we describe and extend the Bayesian framework for concept learning introduced by Tenenbaum (1999). We begin by describing the *generative model*—a joint distribution over concepts and data. The generative model specifies the learner's assumptions about what types of concepts are plausible (the prior) and how concepts give rise to observations (the likelihood). Of central importance is our claim about concept representation: Number concepts are sets generated by a compositional, probabilistic grammar. We then describe how hypothesis sampling can be used to perform approximate inference over number concepts. This sampling-based rational process model provides the basis for our experimental investigations.

Before proceeding, we provide here a non-technical summary of how the model works. A concept is drawn from some space of plausible concepts (the hypothesis space), and examples are drawn from the selected concept. The learner's job is to infer the hidden concept that generated the examples. Because many different concepts can generate any particular set of examples, the problem is fundamentally ill-posed: No single concept is unambiguously "correct". Rather, the optimal inductive inference is a *distribution* over concepts (the posterior), which is computed by multiplying the prior and the likelihood for each potential concept, and then normalizing over the hypothesis space. However, a combinatorial hypothesis space may contain too many hypotheses for complete enumeration to be tractable. A solution to this problem is to randomly sample hypotheses from the posterior and approximate the distribution with a histogram—this is the basis of *Monte Carlo methods* (Robert & Casella, 2004). By limiting the number of samples, a learner can trade off cognitive resources with accuracy: A larger number of samples consumes more cognitive resources (in terms of memory and processing time) while producing a more accurate approximation of the posterior. As we show experimentally, reducing the availability of cognitive resources has deleterious effects on the accuracy of the posterior.

One challenge for practical applications of Monte Carlo methods is that we cannot easily sample from the posterior. To surmount this challenge, we can instead sample from a proposal distribution (e.g., the prior) and then weight the samples to correct for the fact that they were generated from the wrong distribution. When the number of samples is small and the proposal distribution is far from the posterior, this method can lead to degeneracy: a small number of samples have very large weights and the rest of the samples are effectively ignored. This means that the *effective* sample size is smaller than the number of samples. To remedy this problem, we can delete conditionally unlikely samples (i.e., those with small weights) by resampling: generating a new sample set by drawing samples with probability proportional to their weights.

A final challenge is that the examples may arrive sequentially, and it is wasteful to recompute the posterior from scratch after

each new example. Monte Carlo methods can be adapted to this sequential setting by updating the weights online. While making inference more efficient, the combination of sequential updating and resampling can have unwanted side-effects; in particular, if later examples support a hypothesis that was deleted based on earlier examples, the approximation may place zero probability mass on this hypothesis. Thus, sequential hypothesis sampling schemes can give rise to order effects. By looking for such order effects in human behavior, we can place algorithmic constraints on theories of concept learning.

2.1. Generative model

Following Tenenbaum (1999), we assume that participants are shown a set of N positive examples[1] $X = \{x_1, \ldots, x_N\}$ of concept h, and their goal is to compute $P(x' \in h|X)$, the probability that a new observation x' belongs to h given the examples. The target concept h is drawn from a hypothesis space $\mathcal{H}$ with prior probability $P(h)$, described further in the next section.

Under the assumption that examples are sampled uniformly from the set of observations consistent with the concept, Tenenbaum (1999) assigned a likelihood of zero to hypotheses inconsistent with the observed data, but we instead maintain a small probability e^{-b} that the inconsistencies are actually outliers (Goodman et al., 2008; Nelson, Movellan, & Tenenbaum, 2001; Nosofsky et al., 1994). This is equivalent to assuming that an inconsistent hypothesis which has a higher probability under the data than a consistent hypothesis corresponds to a human judgment in which the inconsistency is included as an outlier, e.g. "all multiples of 5, and 17". This leads to the following likelihood:

$$P(X|h) \propto \prod_{n=1}^{N} e^{-b\mathbb{I}\{x_n \notin h\}}, \tag{1}$$

where $\mathbb{I}\{\cdot\} = 1$ if its argument is true, and 0 otherwise.

The probability that a new observation belongs to h given the examples (the generalization function) is given by:

$$P(x' \in h|X) = \sum_{h':x' \in h'} P(h = h'|X), \tag{2}$$

where h' belongs to the subset of hypotheses whose extension includes x'. Tenenbaum and Griffiths (2001) referred to this as *hypothesis averaging*, because it corresponds to averaging the predictions that each hypothesis makes about concept membership, weighted by the posterior probability of the hypothesis, $P(h|X)$. The posterior is calculated using Bayes' rule:

$$P(h|X) = \frac{P(X|h)P(h)}{\sum_{h' \in \mathcal{H}} P(X|h')P(h')}. \tag{3}$$

In the next section, we describe the hypothesis space $\mathcal{H}$ and the prior $P(h)$.

2.2. Concept representation

The framework originally proposed by Tenenbaum (1999) requires enumerating all hypotheses and scoring them using Bayes' rule. Even so, the model captures only a subset of plausible hypotheses: intervals of consecutive numbers and basic mathematical rules. The model of Coen and Gao (2009) used a probabilistic context-free grammar (PCFG) to describe more complex generating functions for ordered sequences. Here we describe a similar approach to representing a combinatorial hypothesis space over unordered sets.

[1] This framework can be extended to negative examples (Goodman et al., 2008), but for simplicity we only deal here with positive examples.

Table 1
Production rules in the grammar. This list contains productions that rewrite the Set operator. The productions that rewrite the AndSet and OrSet operators are the same as for the Set operator, but are assigned different production probabilities in the prior.

Set → AndSet *and* AndSet
Set → OrSet *or* OrSet
Set → Multiples of n
Set → Numbers between n and m
Set → Numbers containing the digit n
Set → Prime numbers
Set → Powers of n
Set → n

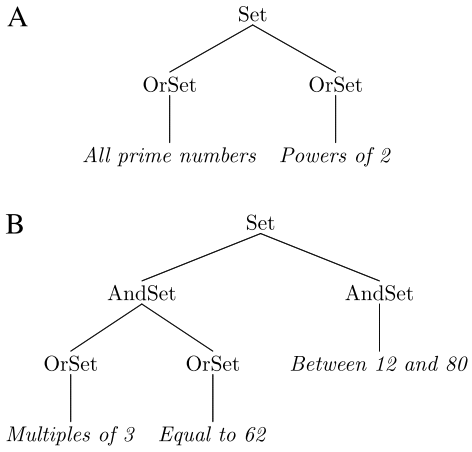

Fig. 1. Example parse trees. (A) Representation of "All prime numbers or powers of 2". (B) Representation of "Numbers between 12 and 80 that are either multiples of 3 or equal to 62".

Our grammar consists of the following elements:

- A set of terminal symbols consisting of functions that take zero or more integer arguments.
- A set of nonterminal symbols. These correspond to three different types of set operations: generation (`Set`), conjunction (`AndSet`), and union (`OrSet`).
- A set of production rules, which map nonterminals to sequences of terminals and nonterminals. The complete set of production rules is shown in Table 1.
- Each production rule is associated with a probability of applying the rule.

A concept is drawn from the hypothesis space by probabilistically composing a sequence of production rules, always starting with the `Set` symbol. Each time a production rule is applied to one of the symbols in the sequence, that symbol is replaced by a sequence of terminals and nonterminals. This process continues until all the nonterminal symbols have been replaced by terminal symbols, resulting in a valid number concept. We can visualize the history of this production process as a parse tree, where the nodes correspond to symbols and edges correspond to applications of production rules. Several examples are shown in Fig. 1.

The prior is parametrized by the production probabilities. The prior over productions can be thought of as three separate priors, for rewriting the Set operator, the AndSet operator, and the OrSet operator. We distribute probability mass within each of these priors by dividing the prior into three subsets:

1. The productions that rewrite to AndSet and OrSet;
2. The productions that rewrite to Interval and Multiples;
3. All remaining productions.

The prior over Set rewrite rules places 1/4 of the probability mass uniformly over the items in subset 1, and 1/4 of the probability mass uniformly over the items in subset 3. The remaining 1/2 of the probability mass is placed on subset 2, weighting Multiples higher than Intervals: specifically, 0.45 probability on Multiples and 0.05 on Intervals.

For both AndSet and OrSet rewrite rules, zero mass is placed on subset 1 and 2/3 of the mass is distributed uniformly over subset 3. This constrains the depth of the generated parse trees, though not enough to make exhaustive enumeration of all hypotheses tractable. The remaining 1/3 of the probability mass is distributed over subset 2 as follows: for AndSet, all of the mass is placed on Multiples and none on Interval; for OrSet, 1/30 of the mass is placed on Multiples and 3/10 of the mass on Interval.

The terminal symbols in the grammar are functions of integer arguments. Each terminal symbol has a corresponding distribution over integers from which its arguments are sampled.

Our prior is hand-tuned, mainly because we have found it difficult to get good fits of the prior parameters to empirical data. However, we have tested this prior on two existing number game data sets, one from (Tenenbaum, 1999), and one from Bigelow and Piantadosi (under review). The correlation between model predictions and human judgments is significantly above zero for both data sets ($p < 0.00001$): $r = 0.78$ for the Tenenbaum data set, and $r = 0.37$ for the Bigelow and Piantadosi data set. Thus, we have independent verification that this prior can predict performance in the number game, although it is also clear that there is still a substantial amount of variance not explained by the prior. We also found that using uniform probabilities in the compositional grammar only changed correlations between model and behavior slightly, indicating that our framework is robust to different assumptions about the prior.

2.3. Hypothesis sampling

Compositional hypothesis spaces will generally have an exponentially large number of summands in the denominator of Eq. (3). As a consequence, exact inference is intractable. Monte Carlo methods can be used to approximate the posterior with a set of M samples (particles), $\{h^1, \ldots, h^M\}$:

$$P(h|X) \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\{h = h^m\}, \tag{4}$$

where $h^m \sim P(h|X)$. The challenge is that in most cases the posterior cannot be sampled from directly. Instead, we can sample from a proposal distribution $Q(h)$ and then weight the samples according to:

$$w^m \propto \frac{P(X|h^m)P(h^m)}{Q(h^m)}. \tag{5}$$

By resampling $\{h^1, \ldots, h^M\}$ with replacement from Multinomial $(w^1, \ldots, w^M)$, we obtain samples approximately distributed according to the posterior. In the limit $M \to \infty$, the approximation converges to the true posterior. This technique is known as *importance sampling*, and the weights are referred to as *importance weights* (Robert & Casella, 2004).

When the examples are observed one at a time, the importance weights can be updated online:

$$w_n^m \propto w_{n-1}^m \frac{P(x_n|h^m)P(h^m)}{Q(h^m)}. \tag{6}$$

Thus, after observing n examples, the weighted samples approximate the posterior $P(h|X)$. This online updating of weights is known as *particle filtering* (Doucet, De Freitas, & Gordon, 2001). We take $Q(h)$ to be the prior $P(h)$, which is easy to sample from. In this case, the weights simplify to $w_n^m \propto P(x_n|h^m)$. In other words, the weights are normalized likelihoods.

It is possible that none of the samples drawn from the prior provides a suitable hypothesis for the data, particularly when the number of samples is small. To deal with this, we introduce a rejuvenation step (Chopin, 2002), which introduces diversity into the sample set by repeatedly applying a Markov transition kernel to each sample. The kernel is chosen so that it leaves the posterior distribution invariant, thereby ensuring that the samples remain correctly distributed. In particular, we perform two iterations of the Metropolis–Hastings algorithm (Robert & Casella, 2004) for each particle, performing a local move by drawing new values for the arguments of the terminal symbols in the parse tree directly from the prior and replacing the existing particle according to the Metropolis–Hastings acceptance rule. Previous research suggests that rejuvenation may play a role in explaining recency effects in causal learning (Abbott & Griffiths, 2011).

3. Garden paths in concept learning

We now turn to the main task of the paper: presenting evidence for hypothesis sampling in human concept learning.[2] As discussed in the Introduction, a key signature of hypothesis sampling is sensitivity to the order of data. If a subset of hypotheses are supported by early evidence, and the rest discarded, then the posterior approximation will show a garden path effect, whereby later evidence favoring the discarded hypotheses is discounted because those hypotheses are no longer in the support of the approximation. To test this prediction, we designed a sequential version of the number game that would allow us to examine order effects. In this version of the number game, numbers are incrementally added to the set, and after each addition participants are asked to judge which numbers are in the concept's extension. Note that there is no sequential structure to the number concepts (i.e., the samples are unordered sets), and participants are instructed to treat each number as an independent draw from the number concept.

We induced garden path effects by manipulating the order of number sequences. We then evaluated order effects by measuring differences in inductive generalization at the end of the sequence. While Tenenbaum's original analysis of the number game (Tenenbaum, 1999) does not predict a difference between conditions, a resource-limited particle filter will tend to be sensitive to order, since there is the possibility of discarding hypotheses during resampling (Abbott & Griffiths, 2011; Levy et al., 2009; Sanborn et al., 2010).[3] To facilitate the detection of order effects, we selected orders that would favor different hypotheses in the early and late phases of the sequence.

If resource limitations arise in part from dividing attention across multiple cognitive processes, then reducing the availability of resources by increasing cognitive load should amplify order effects. We tested this hypothesis by having some participants engage in a secondary distractor task. To simulate the effect of cognitive load, we varied the number of samples available to the particle filter, showing that reducing the number of samples best captures the performance of participants engaged in the distractor task.

[2] Model code and experimental data are available at https://github.com/pratiksha/numbergame.

[3] A rational process model is not the only way to explain order effects. If the generative model captures sequential dependencies, then a purely computational-level analysis will be sensitive to order (Jones, Curran, Mozer, & Wilder, 2013; Navarro, Newell, & Schulze, 2016; Qian & Aslin, 2014; Yu & Cohen, 2009).

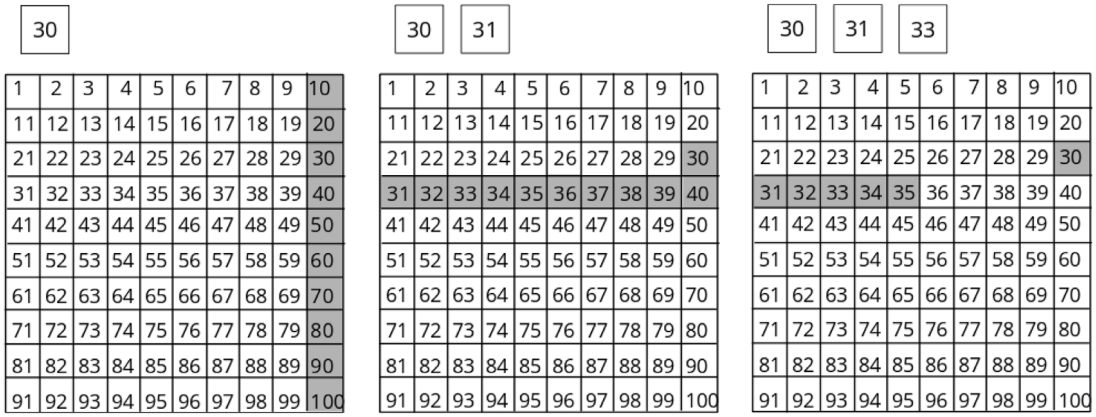

Fig. 2. The experimental setup. On each trial, participants were asked to click on all of the numbers belonging to the concept from which the example numbers were being sampled. Numbers in dark gray on each grid illustrate possible guesses. One example number was revealed on each trial.

3.1. Methods

3.1.1. Participants

One hundred and fifty participants completed the experiments online via the Amazon Mechanical Turk web service. Twenty-seven participants who responded on the final trial with the minimal subset (only numbers in the example set) or maximal subset (all numbers between 1 and 100) were assumed to have misunderstood the experiment instructions, and their data were excluded, leaving 123 participants in subsequent analyses (96 in the first experiment, 27 in the second experiment). Each participant received a payment of $0.50.

3.1.2. Procedure

Participants were shown a sequence of seven numbers derived from a particular subset of numbers between 1 and 100. Numbers in the sequence were revealed one after another. After each number was revealed, participants were asked to identify which numbers between 1 and 100 belonged to the same subset of numbers from which the example set was being generated, by clicking on each number from the hypothesized subset on a grid of numbers. On each trial, participants were required to identify a subset which contained all numbers shown until that trial, and prompted with a warning message if they did not (see Fig. 2).

Participants in the first set of experiments ($N = 96$) were shown one of two orderings of the same set of numbers:

Early: {30, 31, 33, 24, 21, 36, 39}

Late: {30, 33, 24, 21, 36, 31, 39}

The placement of the number 31 earlier in the sequence was intended to bias participants towards the hypothesis "numbers between 20 and 40", while the placement of 31 later in the sequence was intended to elicit a response similar to "multiples of 3, and 31"—that is, we hypothesize that introducing 31 later in the sequence would induce participants to suggest an inconsistency having initially received a large number of positive examples of multiples of 3.

To explore the effects of cognitive load, a subset of participants were asked to simultaneously perform a distractor task. These participants were shown three distractor numbers at the beginning of the task (before the number sequence) and asked to memorize them without writing them down. At the end of the number sequence, they were asked to supply the three numbers they memorized.

To assess the generality of our experimental manipulation and model predictions, we ran another set of participants ($N = 27$) on another sequence with a larger range of numbers:

Early: {30, 31, 45, 24, 6, 12, 60}

Late: {30, 12, 45, 24, 6, 31, 60}

Similarly to the other number sequences, the placement of the number 31 earlier in the sequence was intended to bias participants towards the hypothesis "numbers between 0 and 60", while the placement of 31 later in the sequence was intended to elicit a response similar to "multiples of 3, and 31".

3.2. Results and discussion

To summarize the dynamics of participants' posterior over number concepts, we computed, for each trial, the absolute deviation (absolute difference, summed over all numbers) between the average judgment and two number concepts: "multiples of 3, and 31" and "numbers between 20 and 40". An ideal Bayesian learner will place most of the posterior probability mass on this concept. As shown in Fig. 3, participants in the Early condition exhibited a steady decline in the absolute deviation from the interval concept, indicating that they gradually converged to the ideal inference. In contrast, participants in the Late condition exhibited a steady increase in the absolute deviation from the interval concept, indicating that their posterior was ambling down a garden path. The deviation from the multiples concept shows the opposite pattern: Participants in the Late condition were drawn towards the multiples concept. Only after the number 31 appeared in the sequence did the deviation from the interval concept drop and the deviation from the multiples concept increase. However, the deviation never changed to the same level as in the Early condition. Thus, participants in the Late condition were not able to fully recover from the early information in the sequence, consistent with the idea that they tended to discard the interval hypothesis. This pattern is captured by our model (Fig. 4; see below for details on how model predictions were obtained).

Fig. 3 also shows that the order effect (difference between Early and Late) was amplified when participants performed the distractor task. This is consistent with the idea that performing the distractor task depleted cognitive resources and effectively reduced the number of samples. To compare the strength of the order effect in the Distractor and No distractor conditions, we calculated the summed absolute difference between the final judgments for the Early and Late conditions, where the sum was taken over all possible numbers, averaged across participants within each condition (Fig. 5). Using bootstrapped confidence interval estimation, we found that the order effect was significantly stronger in the Distractor condition ($p < 0.05$). This result was

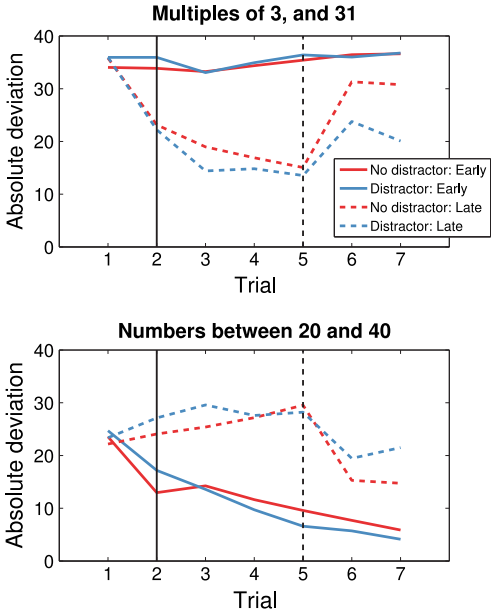

Fig. 3. Deviation between average human judgments and two number concepts. Smaller values on the Y-axis indicate that the inferred concept is closer to either "multiples of 3, and 31" (top) or "numbers between 20 and 40" (bottom). The vertical lines indicate the trial on which the number 31 appears for the Early conditions (solid line) and Late conditions (dashed line). The Early sequence consists of [30, 31, 33, 24, 21, 36, 39] and the Late sequence consists of [30, 33, 24, 21, 36, 31, 39].

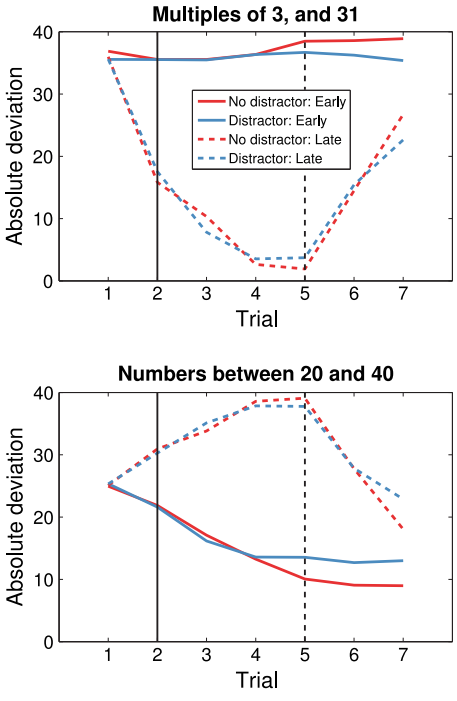

Fig. 4. Deviation between model predictions and two number concepts. Smaller values on the Y-axis indicate that the inferred concept is closer to either "multiples of 3, and 31" (top) or "numbers between 20 and 40" (bottom). The vertical lines indicate the trial on which the number 31 appears for the Early conditions (solid line) and Late conditions (dashed line). The Early sequence consists of [30, 31, 33, 24, 21, 36, 39] and the Late sequence consists of [30, 33, 24, 21, 36, 31, 39].

captured by the particle filter model, where we assumed that the model used 130 particles in the No distractor conditions and 70 particles in the Distractor conditions (see below for justification of this modeling choice).

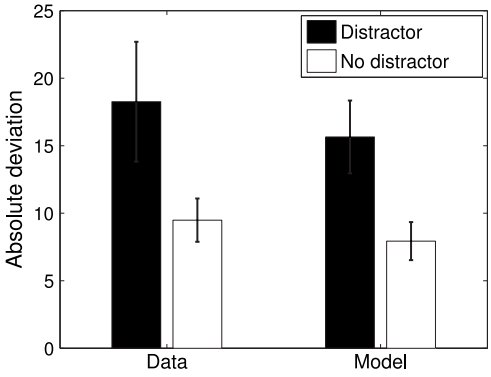

Fig. 5. Quantification of order effects. The Y-axis shows the summed absolute difference between the final judgments for the Early and Late conditions, where the sum was taken over all possible numbers, averaged across participants within each condition. Larger values indicate a stronger order effect. Error bars represent 95% bootstrapped confidence intervals. For both the data and the model, the order effect is significantly stronger in the Distractor condition compared to the No distractor condition ($p < 0.05$). The Early sequence consists of [30, 31, 33, 24, 21, 36, 39] and the Late sequence consists of [30, 33, 24, 21, 36, 31, 39].

To present a finer-grained picture of the experimental results, we plotted the human judgments for the final trial in each of the four conditions. As shown in Fig. 6, participants in the Early condition tend towards the interval hypothesis, but participants in the Late condition show a tendency towards the "multiples of 3, and 31" hypothesis. This effect was accentuated when participants performed the task under cognitive load. The particle filter model reproduces this pattern, although overall it underestimates the size of the order effect (Fig. 7). We suspect that this underestimation occurs because humans are using a more data-driven proposal mechanism that induces stronger order effects. In other words, if participants use the data to heuristically generate guesses, they will be more strongly anchored to the initial data compared to if they sample from their prior, as in our model.

For instance, one function that the rejuvenation step serves is to allow the model to generalize to larger intervals as more data points are observed, by resampling interval hypotheses; however, this step leads to intervals being overrepresented in the posterior relative to human judgments. This generalization might be better explained by, for instance, a more complex mechanism that dynamically adapts the parameters of existing particles.

Similarly, two choices made in designing the prior (as described in Section 2.2) might be accounted for by a more sophisticated grammar or sampling procedure: first, limiting the depth of the hypotheses, without which the probability of overfitting to the data with a complicated hypothesis is too high; and second, placing more probability mass on mathematical hypotheses than on intervals, without which the model overrepresents intervals.

To obtain model predictions, we computed the maximum *a posteriori* (MAP) hypothesis for each run of the particle filter, and averaged this MAP hypothesis over many runs. This corresponds to the assumption that participants are reporting the MAP hypothesis, and we are trying to capture the aggregate behavior across participants. We also assumed that participants in the Distractor conditions had fewer particles available than participants in the No distractor conditions. We arrived at this assumption quantitatively by computing the correlation between model predictions and human judgments on the final trial for different numbers of particles. Overall, the correlations are relatively high (Fig. 8), peaking at 70 particles for the Distractor condition and peaking at 130 particles for the No distractor condition. The same number of particles was favored in each condition when we computed the correlation over all trials instead of just the last one. Thus, our experimental data are well explained

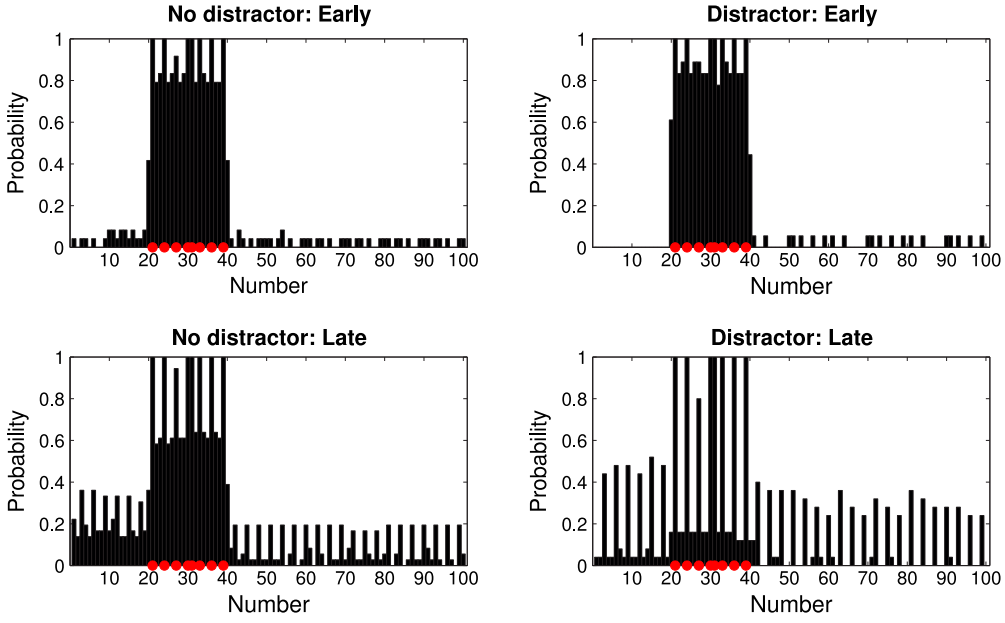

Fig. 6. Human data. Generalization probabilities for the final trial. Red circles indicate the set of numbers shown to participants (note that these are the same in all panels). The Early sequence consists of [30, 31, 33, 24, 21, 36, 39] and the Late sequence consists of [30, 33, 24, 21, 36, 31, 39]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

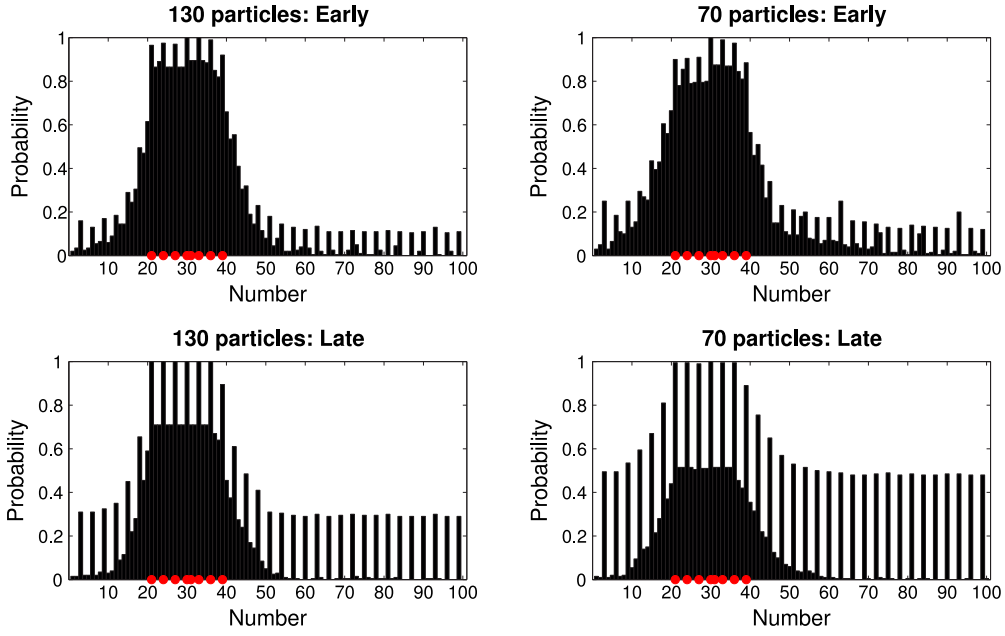

Fig. 7. Model predictions. Generalization probabilities for the final trial. Red circles indicate the set of numbers shown to participants (note that these are the same in all panels). The Early sequence consists of [30, 31, 33, 24, 21, 36, 39] and the Late sequence consists of [30, 33, 24, 21, 36, 31, 39]. The left panels show the results of running a particle filter with 130 particles, and the right panels show the results with 70 particles to simulate the effects of cognitive load induced by the distractor task. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by a resource-constrained hypothesis sampling algorithm that is sensitive to cognitive load (see also Vul et al., 2009).

To explore the effects of prior hyperparameters (namely, the production probabilities), Fig. 8 also shows the correlations computed using uniform probabilities for all productions. While the correlations are still fairly high, the peak correlations for both conditions are both lower using uniform predictions (0.86 for Distractor and 0.92 for No Distractor) compared to tuned probabilities (0.90 for Distractor and 0.93 for No Distractor). Thus, the tuned probabilities have a slight advantage in fitting our data.

There is one clear discrepancy between the model and data: the model does not capture the sharp change in generalization probabilities outside the minimal subset. Instead, the model shows a graded falloff, because it is averaging over many interval hypotheses. This suggests that our parametrization of the prior does not yet fully capture human inductive biases about number concepts.

Finally, we tested the generality of our model predictions, without any additional parameter tuning, by collecting human data on a sequence with a wider range of numbers (see Methods). The results, shown in Fig. 9, reveal a clear order effect, replicating our earlier finding with a narrower range of numbers. Although the model again does not capture the sharp change in generalization probabilities outside the minimal subset, it still reproduces a

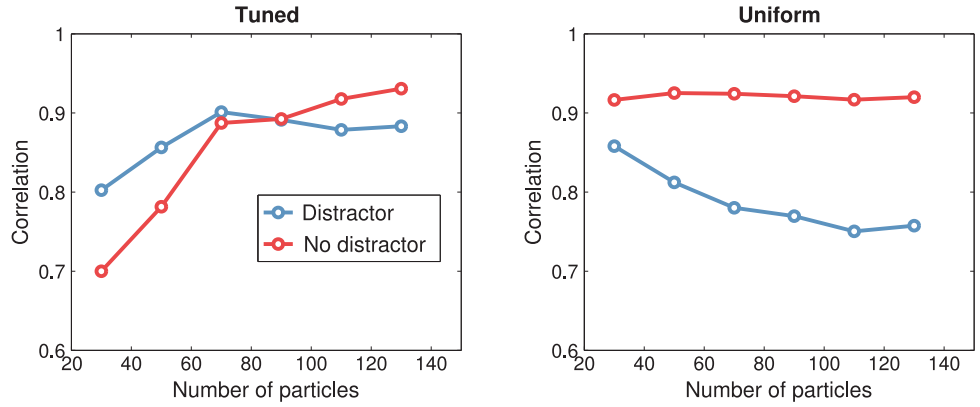

Fig. 8. Correlation between model predictions and human judgments. Pearson correlations for Distractor and No distractor conditions are averaged across Early and Late conditions. (Left) Model predictions with tuned prior probabilities. (Right) Model predictions with uniform prior probabilities.

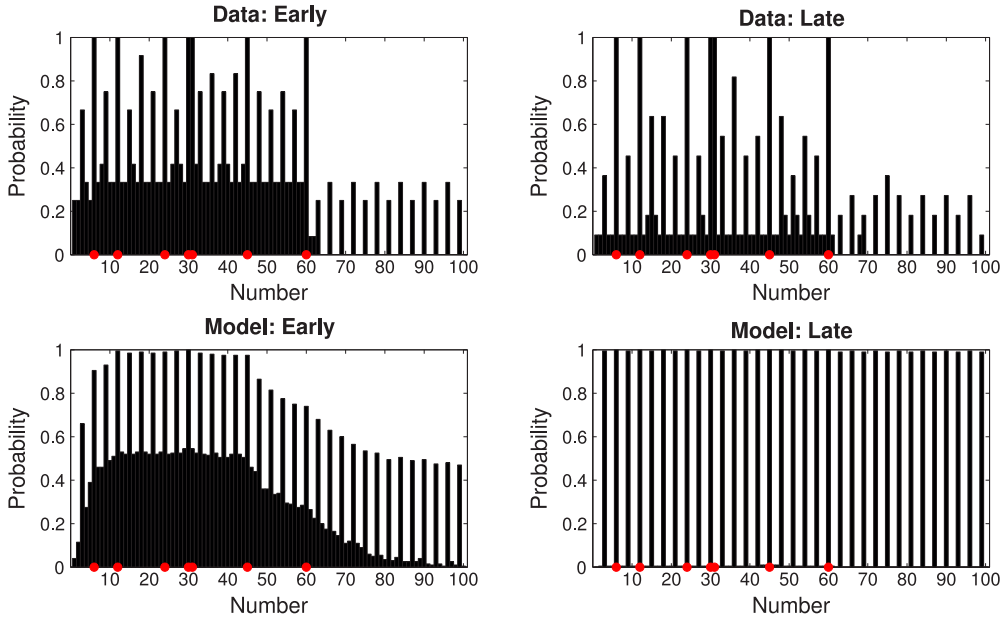

Fig. 9. Human and model generalization probabilities on the final trial with a larger range of numbers. Red circles indicate the set of numbers shown to participants (note that these are the same in all panels). The Early sequence consists of [30, 31, 45, 24, 6, 12, 60] and the Late sequence consists of [30, 12, 45, 24, 6, 31, 60]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

strong order effect, consistent with the human data. The model using 130 particles provides a good quantitative match, with a correlation of $r = 0.81$ averaged across conditions.

4. Discussion

Our experiment provides support for hypothesis sampling as a theory of inference in compositional hypothesis spaces. We used a probabilistic grammar as the prior in our rational analysis of number concept learning, extending earlier work that used a simpler, non-compositional hypothesis space (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001). We then leveraged this grammar to make quantitative predictions about concept learning in a novel sequential version of the number game. Participants showed clear garden path effects that were sensitive to stimulus order. In particular, we found that participants failed to infer a high probability hypothesis when that hypothesis was disfavored by early evidence, consistent with the idea that such hypotheses are discarded by the particle filtering algorithm. We further showed that this order effect could be strengthened by placing participants under cognitive load, as though the reduction in resources forced them to use fewer particles.

While it is unlikely that our model is correct in its details (both the hypothesis space and the sampling algorithm are simple compared to human capabilities), our data nonetheless provide constraints on the kind of cognitive architecture that could reproduce both the successes and failures of human symbolic concept learning. At present, hypothesis sampling provides the only plausible process-level account of how humans could successfully explore complex hypothesis spaces (Ullman et al., 2012), and this view is supported by our finding that human inferences about number concepts are strongly correlated with the predictions of our hypothesis sampling model. The distinctive failures of human learning, such as garden path effects, arise from the resource-limited nature of the sampling process (Abbott & Griffiths, 2011; Levy et al., 2009). The intersection of complex hypothesis spaces and approximate inference provides a parsimonious account of our findings. The main contribution of our paper is a detailed empirical and theoretical study of this intersection.

Our finding that cognitive load has a deleterious effect on hypothesis sampling resonates with earlier work on the role of working memory in probability judgment. Dougherty and Hunter found that the number of hypotheses participants generated was correlated with a measure of working memory capacity (Dougherty &

Hunter, 2003a,b). Furthermore, participants generated fewer hypotheses when they were put under time pressure (Dougherty & Hunter, 2003b) or cognitive load (Sprenger et al., 2011). These findings provide converging evidence for our assertion that garden path effects and other frailties of probabilistic reasoning arise from cognitive resource limitations.

A sequential variant of the number game similar to the one presented in this paper has been previously explored by Austerweil and Griffiths (2011) and Coen and Gao (2009). Both studies contain similarities to our work, particularly in the representation of the hypothesis space as a grammar composed of rules and the derivation of the prior from human data. An important difference is that these studies addressed the learning of *sequence* (i.e., ordered set) concepts, whereas our work addresses the learning of *unordered set* concepts. Our order effects thus likely reflect inferential processes rather than perceptions of sequential structure. We pursued this idea by introducing a cognitive load manipulation that ostensibly affected these inferential processes.

We now turn to a broader discussion of empirical evidence in support of hypothesis sampling theories, and then discuss several future directions for this research.

4.1. Evidence for hypothesis sampling in human cognition

The evidence for hypothesis sampling can largely be divided into two classes: *response variability* and *temporal dynamics*. We will discuss each of these in turn.

Essentially all mechanistic theories of cognition make some provision for response variability. Some theories view variability as the result of irreducible "noise" (e.g., due to the stochasticity of neural firing; Faisal, Selen, & Wolpert, 2008) or an illusion based on scientific ignorance (Skinner, 1974), whereas other theories ascribe a functional role to variability. Variability can contribute to exploration strategies in reinforcement learning (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), increase fitness across a population of foragers (Kamil & Roitblat, 1985), and enable an agent to behave unpredictably in a competitive game (Glimcher, 2005; Smith, 1982). An important constraint on theoretical interpretations of variability is the "matching law" (Herrnstein, 1970): The probability of selecting a response is proportional to the probability that the response is correct. While numerous reinforcement learning theories have been formulated to account for matching behavior, these theories leave unexplained why matching occurs in high-level cognitive tasks. For example, Denison et al. (2013) showed that children, when asked to provide multiple guesses in a causal inference task, distribute hypotheses according to their posterior probability, in accordance with Bayesian hypothesis sampling. Studies with adults have shown similar results across a variety of tasks (Goodman et al., 2008; Moreno-Bote et al., 2011; Vul et al., 2014; Wozny et al., 2010).

Most Bayesian hypothesis sampling theories will (at least roughly) predict probability matching behavior. However, the theories may differ in their predictions about temporal dynamics. Goodman et al. (2008) suggested that participants sample one concept at a time according to a Markov chain, thus implementing a form of MCMC. A similar mechanism has been invoked in other domains (e.g., Gershman et al., 2012; Lieder, Griffiths, & Goodman, 2012; Ullman et al., 2012), and relates to an older literature on serial hypothesis testing (Brown, 1974). In contrast, the particle filter implies that participants represent an ensemble of hypotheses at each moment. A key prediction of the MCMC account is that the sampling process will give rise to autocorrelation of hypotheses (but see Bonawitz, Denison, Gopnik, & Griffiths, 2014, for an alternative viewpoint on hypothesis autocorrelation). Consistent with this prediction, hypotheses tend to be "anchored" to initial guesses (Lieder et al., 2012), and increasing the time between eliciting hypotheses decreases their autocorrelation (Denison et al., 2013; Vul & Pashler, 2008). Another distinctive characteristic of MCMC is its rich internal dynamics; Gershman et al. (2012) showed how MCMC could produce perceptual fluctuations like multistability and traveling waves in binocular rivalry experiments.

Another important difference is that MCMC operates over the entire data set, whereas particle filtering operates online (one data point at a time). Online inference appears to be important for explaining order effects (e.g., Levy et al., 2009; Sanborn et al., 2010), since algorithms that operate over the entire data set should be invariant to order once all data points have been observed.

4.2. Future directions

An important question for future research is what sort of sampling algorithm best describes concept learning behavior. For the concept learning experiments described in this paper, an online learning algorithm like particle filtering seems *a priori* more cognitively plausible than MCMC, and naturally leads to order effects such as garden paths. Our implementation actually employs a combination of particle filtering and MCMC (via the rejuvenation step; see also Abbott & Griffiths, 2011). More work is needed to tease apart the precise contributions of these different sampling mechanisms.

Another challenge for our computational framework is understanding the hypothesis generation process. We proposed that hypotheses are sampled from the prior (see also Shi, Griffiths, Feldman, & Sanborn, 2010); while there is some evidence that hypotheses with high prior probability are preferentially sampled (Dougherty & Hunter, 2003a; Weber, Böckenholt, Hilton, & Wallace, 1993), there is also reason to think that the hypothesis sampling process is more data-driven (Cherubini, Castelvecchio, & Cherubini, 2005; Lewis, Perez, & Tenenbaum, 2014; Schulz, 2012). For example, the phenomenon of *base-rate neglect* (Bar-Hillel, 1980) suggests that in some cases participants discount their priors, relying instead on the likelihood (i.e., the match between hypothesis and data). In a related vein, Schulz (2012) has argued that hypothesis sampling is sensitive to discrepancies between prediction and observation: Learners postulate hypotheses that fix specific errors in their current crop of hypotheses. There is a rich literature on hypothesis generation (e.g., Dougherty & Hunter, 2003a; Gettys & Fisher, 1979; Mehle, 1982; Thomas, Dougherty, Sprenger, & Harbison, 2008), but Bayesian hypothesis sampling theories have made little contact with this literature (for exceptions, see Bonawitz & Griffiths, 2010; Navarro & Perfors, 2011).

Finally, some of the most exciting work in concept learning has studied tasks in which participants can actively make choices to gather data (e.g., Navarro & Perfors, 2011; Nelson et al., 2001; Tsividis, Gershman, Tenenbaum, & Schulz, 2013). These active learning tasks raise a host of questions concerning information gain, confirmation bias, and exploration–exploitation trade-offs. Rational process models of inference can potentially make unique predictions about cognitive load and order manipulations in such tasks.

4.3. Conclusions

We began this paper with the puzzle of how humans can efficiently make inferences about complex concepts. The answer pursued here is that humans use an approximate inference algorithm (particle filtering) that explores compositional hypothesis spaces through sampling (Griffiths et al., 2012). Particle filtering accounts for both the strengths and weaknesses of human concept learning. On the one hand, it explains how humans are able to acquire richly structured concepts like numbers. On the

other hand, particle filtering also explains the failure modes of concept acquisition—the deleterious effects of garden paths and cognitive load. These failure modes arise from a form of "bounded rationality" (Simon, 1982), whereby computational costs and statistical accuracy are traded off against one another to optimize the performance of the system (Gershman, Horvitz, & Tenenbaum, 2015; Griffiths, Lieder, & Goodman, 2015; Vul et al., 2014). By exploring both of these aspects, our work offers insight into how rational analysis connects with cognitive mechanisms.

Acknowledgments

We thank Ed Vul for helpful discussions, Ameesh Goyal and Kenneth Siebert for providing code for an initial version of the experimental interface, and Steve Piantadosi and Eric Bigelow for providing us with their data. This research was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL), under contract FA8650-14-C-7358. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the US Government. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Abbott, J.T., & Griffiths, T.L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the 33rd annual conference of the cognitive science society*.

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*, 499–526.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Bonawitz, E., & Griffiths, T.L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 2260–2265).

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

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

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 Section A*, *58*, 309–332.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, *89*, 539–552.

Coen, M., & Gao, Y. (2009). Learning from games: inductive bias and Bayesian inference. In *Proceedings of the 31st annual conference of the cognitive science society*.

Daw, N. D., & Courville, A. C. (2007). The pigeon as particle filter. *Advances in Neural Information Processing Systems*, *20*, 369–376.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in childrens causal inferences: The sampling hypothesis. *Cognition*, *126*, 285–300.

Doucet, A., De Freitas, N., Gordon, N., et al. (2001). *Sequential Monte Carlo methods in practice*. New York: Springer.

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

Dougherty, M. R., & Hunter, J. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, *31*, 968–982.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*, 292–303.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*, 273–278.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*, 1–24.

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

Glimcher, P. W. (2005). Indeterminacy in brain and behavior. *Annual Review of Psychology*, *56*, 25–56.

Goodwin, G. P., & Johnson-Laird, P. (2011). Mental models of Boolean concepts. *Cognitive Psychology*, *63*, 34–59.

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

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*, 243–266.

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review*, *120*, 628–666.

Kamil, A. C., & Roitblat, H. L. (1985). The ecology of foraging behavior: implications for animal learning and memory. *Annual Review of Psychology*, *36*, 141–169.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*, 685–722.

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

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in Neural Information Processing Systems*, *21*.

Lewis, O., Perez, S., & Tenenbaum, J. (2014). Error-driven stochastic search for theories and concepts. In *Proceedings of the 36th annual conference of the cognitive science society*.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems* (pp. 2690–2798).

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*, 157–201.

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

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*, 12491–12496.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*, 120–134.

Nelson, J. D., Movellan, J. R., & Tenenbaum, J. B. (2001). Active inference in concept learning. In *Advances in neural information processing systems 13*.

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

Piantadosi, S.T., Tenenbaum, J.B., & Goodman, N.D. (2010). Beyond Boolean logic: exploring representation languages for learning complex concepts. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 859–864).

Qian, T., & Aslin, R. N. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences*, *111*, 14400–14405.

Robert, C. P., & Casella, G. (2004). *Monte carlo statistical methods*. New York: Springer.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Schulz, L. (2012). Finding new facts; thinking new thoughts. *Advances in Child Development and Behavior*, *43*, 269–289.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*, 443–464.

Simon, H. A. (1982). *Models of bounded rationality: Empirically grounded economic reason*. MIT Press.

Skinner, B. F. (1974). *About behaviorism*. Knopf.

Smith, J. M. (1982). *Evolution and the theory of games*. Cambridge University Press.

Sprenger, A. M., Dougherty, M. R., Atkins, S. M., Franco-Watkins, A. M., Thomas, R. P., Lange, N., & Abbs, B. (2011). Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Psychology*, *2*.

Tenenbaum, J. B. (1999). A Bayesian framework for concept learning. *Massachusetts Institute of Technology*,.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155–185.

Tsividis, P., Gershman, S.J., Tenenbaum, J.B., & Schulz, L. (2013). Information selection in noisy environments with large action spaces. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1622–1627).

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*, 455–480.

Vul, E., Frank, M. C., Alvarez, G., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems*, *22*, 1955–1963.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, *19*, 645–647.

Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1151–1164.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology*, *6*, e1000871.

Yi, M. S., Steyvers, M., & Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *The Journal of Problem Solving*, *2*, 5.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior? In *Advances in neural information processing systems* (pp. 1873–1880).