Probability Without Counting and Dividing: A Fresh Computational Perspective

Thomas R. Shultz[23] & Ardavan S. Nobandegani[13]
thomas.shultz@mcgill.ca, ardavan.salehinobandegani@mail.mcgill.ca
[1]Department of Electrical & Computer Engineering, McGill University
[2]School of Computer Science, McGill University
[3]Department of Psychology, McGill University

Abstract

Recent experiments show that preverbal infants can reason probabilistically. This raises a deep puzzle because infants lack the counting and dividing abilities presumably required to compute probabilities. In the standard way of computing probabilities, they would have to count or accurately estimate large frequencies and divide those values by their total. Here, we present a novel neural-network model that learns and uses probability distributions without explicit counting or dividing. Probability distributions emerge naturally from neural-network learning of event sequences, providing a computationally sufficient explanation of how infants could succeed at probabilistic reasoning. Several alternative explanations are discussed and ruled out. Our work bears on several other active literatures, and it suggests an effective way to integrate Bayesian and neural-network approaches to cognition.

Keywords: infants; probabilistic learning and inference; neural networks; sampling; control for frequency

Introduction

Succeeding in an uncertain world requires probabilistic reasoning: the ability to compute and act on relevant probabilities. Classical studies found that probabilistic reasoning does not develop in humans until around seven years of age (Piaget & Inhelder, 1975). However, a series of recent experiments shows that preverbal human infants can learn and reason with probabilities (Denison, Reed, & Xu, 2013; Teglas et al., 2011; Xu & Garcia, 2008), even using them to guide their behavior (Denison & Xu, 2010, 2014).

Collectively, the infant experiments present a deep and largely unnoticed puzzle: How could preverbal infants compute probability without explicitly counting and dividing? The standard method of calculating event probabilities is to divide the frequencies of each event by the number of possible events (Kolmogorov, 1956; Moran, 1968). Here, we propose and evaluate a way to learn probability distributions with a computational system called Neural Probability Learning and Sampling (NPLS) that does not require explicit counting and dividing. Our results accurately simulate the infant data and generate testable predictions. We relate our work to several other active research lines on probabilistic and quantitative reasoning.

Four of these experiments with 10-12-month-old infants ruled out a possible alternative explanation that infants might be using raw frequencies rather than probabilities (Denison & Xu, 2014). Earlier infant experiments invariably had confounded probabilities with raw event frequencies (Denison et al., 2013; Denison & Xu, 2010; Teglas et al., 2011; Xu & Garcia, 2008). This confounding of probability and frequency raised the possibility that infants might be using event frequencies rather than computing probabilities.

In the unconfounded experiments (Denison & Xu, 2014), infants were first exposed to two lollipop-like objects of two different colors in live displays and encouraged to approach the one they preferred (see Figure 1). Most approached the attractive pink one rather than the plain black one. Then infants saw two jars containing different proportions of these object colors. The jars were then covered, and one object was randomly removed from each jar and hidden in a separate cup, without revealing its color as only the lollipop handle was visible. When invited to approach and get the item they wanted, would they approach the cup with content drawn from the jar that held more of the preferred color or the other cup with content that was drawn from the jar with a higher proportion of the preferred color? Infant searches closely approximated the favorable proportions, indicating accurate learning and use of the probability distributions.

In Experiment 1, the more favorable jar had a 12:4 preferred item to un-preferred item ratio, while the unfavorable jar had a ratio of 12:36, yielding ground-truth probabilities of .75 vs. .25 (Table 1). Experiment 2 pitted probabilities against frequencies, with ratios of 16:4 vs. 24:96, yielding ground-truth probabilities of .8 vs. .2. Experiment 3 was designed to test whether infants used a different heuristic, raw frequencies of un-preferred objects rather than proportions of preferred objects, using ratios of 8:14 vs. 8:72, yielding unnormalized ground-truth probabilities of .36 vs. .1, respectively. Finally, Experiment 4 challenged infants to distinguish a more subtle probability difference: .8 vs. .6. This was implemented with ratios of 6:15 vs. 60:40. In each of the four experiments, infant search pattern proportions (.75, .79, .75, and .71 for Experiments 1-4, respectively) closely matched normalized ground-truth probabilities (.75, .8, .78, and .57 for Experiments 1-4, respectively). The ground-truth binary search probabilities for Experiments 3 and 4 are calculated by dividing the unnormalized favorable and unfavorable ground-truth probabilities by their respective sums, as the probabilities of mutually exclusive and exhaustive events must sum to 1.

Here, we simulate these four infant experiments with an enhanced neural-network model that also successfully simulates eight other infant probabilistic reasoning experiments reported in four other articles (Denison et al., 2013; Denison & Xu, 2010; Teglas et al., 2011; Xu & Garcia, 2008), thus covering most of the research on preverbal infant learning and use of probability distributions. Due to lack of space, these additional simulations are omitted here.

Table 1: Ratios and probabilities in the four experiments

Experiment	1	2	3	4
Favorable ratio	12:4	16:4	8:14	60:15
Unfavorable ratio	12:36	24:96	8:72	60:40
Favorable probability	.75	.80	.36	.80
Unfavorable probability	.25	.20	.10	.60

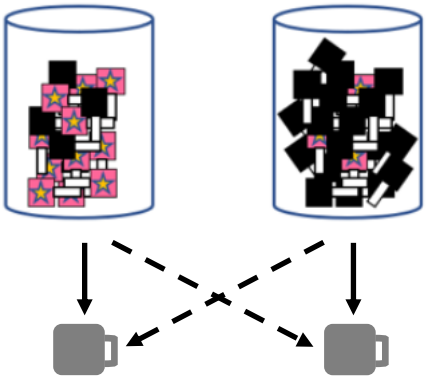

Figure 1: Schematic representation of a test trial in the infant experiments. Infants were first familiarized with two populations with varying ratios of preferred vs. un-preferred objects. Next, the jars were covered and the experimenter randomly removed an object from each jar and placed it in one of the cups, either in front of the source jar for half the infants (solid lines) or the other jar for the other half of the infants (dashed lines). Then, each infant was invited to approach the cups to get their preferred object. Adapted from Denison and Xu (2014).

Methods

Learning in NPLS is based on a constructive neural learning algorithm called SDCC (Sibling-Descendant Cascade-Correlation) that builds the interior of a neural network during learning, and has simulated many deterministic developmental phenomena in infants (Shultz, 2010, 2017; Shultz & Cohen, 2004). Probability distributions are estimated by a network's output activation.

SDCC networks are deterministic, feed-forward, networks that learn from examples by reducing overall prediction error (Baluja & Fahlman, 1994). Unit activations are passed forward from input units that describe examples to hidden units that transform inputs into more abstract representations, and finally to output units coding the response to particular input. Network output can be considered an expectation of what will happen at the output, while target output represents what is actually observed. During learning (in output phase), connection weights are adjusted to reduce network error:

$$E = \sum_o \sum_p \left(A_{op} - T_{op}\right)^2 \qquad (1)$$

where E is sum-of-squared error, A is the actual output activation for unit o and pattern p, and T is the target output activation for this unit and pattern.

SDCC learning starts with a two-layer network (i.e., only the input and the output layer), and then recruits hidden units one at a time to solve the problem being learned. The algorithm constructs its own network topology, as opposed to being designed by a programmer. In input phase, input weights to candidate hidden units are trained to increase the covariation of candidate hidden unit output activation with network error. The highest correlating unit is then installed either on the highest layer of hidden units or on its own higher layer, depending on which has the better absolute covariation with network error. Input weights to each recruited hidden unit are frozen when the unit is installed. Weights are adjusted only one layer at a time, thus never requiring propagation of error signals backwards through the network. The function to maximize in input phase is a covariance between candidate-hidden-unit activation and network error:

$$C = \frac{\sum_o \left| \sum_p (h_p - \langle h \rangle)(e_{op} - \langle e_o \rangle) \right|}{\sum_o \sum_p \left(e_{op} - \langle e_o \rangle\right)^2} \qquad (2)$$

where h_p is activation of the candidate hidden unit for pattern p, $\langle h \rangle$ is the mean activation of the candidate hidden unit for all patterns, e_{op} is the residual error at output o for pattern p, and $\langle e_o \rangle$ is the mean residual error at output o for all the training patterns.

The networks use an asymmetric sigmoid activation function:

$$y_i = \frac{1}{1 + e^{-x_i}} \qquad (3)$$

where y is the output of receiving unit i, x is the net input to unit i, and e is the exponential function. Output activation thus ranges from 0 to 1, just like probabilities do.

Several enhancements of SDCC are required to cope with learning and using probability distributions. First, because of its determinism, SDCC was not satisfied with the high error of probabilistic outcomes, recruiting new hidden units *ad infinitum*. This problem was solved by allowing NPLS to track its progress in error reduction over learning cycles. SDCC already had the capacity to monitor progress within both input and output phases, using parameters for threshold and patience. In output phase, SDCC adjusts connection weights to reduce error. When error reduction stagnates, the algorithm changes to input phase to recruit a new hidden unit, adjusting weights entering candidate units to increase the correlation between their activations and network error. In each of these two phases, stagnation is detected when there is no progress greater than the threshold parameter for the number of training epochs specified by the patience parameter. We extended this scheme by adding an outer loop with its own threshold and patience parameters to monitor

progress over learning cycles, where each such cycle is an input phase and the next output phase (Shultz & Doty, 2014). This allows NPLS to stop when learning stagnates. With this ability, NPLS can learn any unnormalized multivariate probability distribution from examples that specify whether or not an output occurs in the presence of a particular input (Kharratzadeh & Shultz, 2016).

We run 20 NPLS networks in each of the four infant experiments, training them on event sequences with an input unit arbitrarily coding for the source jar (1 or 2) and an output unit coding 1 for presence and 0 for absence of an object type. With this deterministic binary coding, corresponding directly to the visual stimuli presented to the infants, networks learn to output the probability of drawing a preferred item from a favorable and an unfavorable source. Note that ground-truth probabilities are not used as learning targets; they are instead an emergent property of NPLS learning.

Table 2 shows an example of the coding scheme for a simple binary distribution with ratios of 3:1 vs. 1:3. This requires 4 training patterns for each ratio. In 3 of 4 examples for container 1, a focal object appears. For container 2, a focal object appears in only 1 of 4 examples. Our simulations use the exact ratios from the infant experiments, realistically representing what the infant sees in the jars. There is an asymmetric sigmoid activation function on the output unit to keep outputs in 0-1 range of probabilities.

Table 2: Schematic coding of a binary probability distribution

3:1		1:3	
Input	Output	Input	Output
1	1	2	1
1	1	2	0
1	1	2	0
1	0	2	0

A second problem is that SDCC could not probabilistically generate novel examples from example categories that it had learned. Following recent advances (Nobandegani & Shultz, 2017), we pair a Markov-chain Monte Carlo sampling algorithm (MCMC) with each network to simulate how infants generate the more favorable container from the category of the preferred object, thus converting a deterministic neural network into a probabilistic generative model.

Infant selection patterns can be mathematically characterized as a form of sampling from the underlying probability distribution. An infant could mentally draw a sample, cued by desire for the high probability of a preferred object, and this would identity the more favorable jar for obtaining that object. This sampling could guide physical search towards the cup supplied by that favorable jar.

NPLS induces a probability distribution $p(\mathbf{X}|\mathbf{Y})$ on the deterministic input-output mapping $f(\mathbf{X}; W^*)$ learned by an NPLS network, and uses MCMC to sample from that induced distribution. The induced distribution is given by:

$$p(\mathbf{X}|\mathbf{Y} = Y) \propto exp(-\beta||Y - f(\mathbf{X}; W^*)||_2^2) \qquad (4)$$

where $||\cdot||_2$ is the l_2-norm, W^* the set of weights for a network after training, and β a damping factor. For an input instance $\boldsymbol{X} = X$ belonging to the desired class Y, the network output $f(X; W^*)$ is expected to be close to Y in the l_2-norm sense. Equation 4 adjusts the probability of input instance X to be inversely proportional to the base-e exponentiation of the l_2 distance. Our NPLS system can handle any MCMC method, including Metropolis-Adjusted Langevin, a gradient-based MCMC method, which can be implemented in a biologically-plausible way (Moreno-Bote, Knill, & Pouget, 2011; Savin & Denève, 2014).

Results

For each experiment, the mean network probability estimates closely match ground-truth probabilities, consistent with the hypothesis that the infants were computing relevant probabilities (Denison & Xu, 2014). The mean estimated probability for the favorable location is considerably higher than that for the unfavorable location in every simulated experiment, as tested with a paired-sample t-test, $p < .0001$. All p values in this article represent 2-tailed comparisons. Mean network output activations correlate highly with ground-truth probabilities across the eight conditions of the four experiments, $r(6) = 1.0$, $p = 9.1E\text{-}11$.

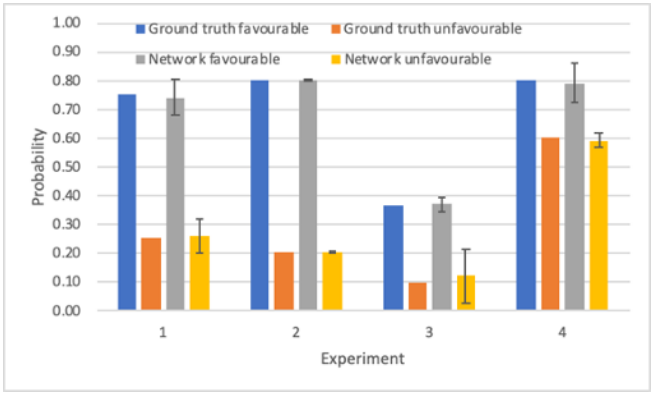

Figure 2. Ground-truth (blue and orange) and mean simulated probabilities (grey and gold) for each of four experiments on infant probabilistic reasoning (Denison & Xu, 2014). Error bars are standard deviations for 20 networks.

Figure 2 shows the matches between of ground-truth probabilities and mean network output, with standard deviation bars around the simulation means. In each experiment, the size and direction of the location difference between favorable and un-favorable activations is apparent (grey vs. gold) and very close to ground truth probabilities (blue vs. orange). This represents the probabilistic knowledge that enables the sampling that could guide infant crawling towards the more favorable location for their preferred object. In contrast, predictions based on relative frequencies of the preferred object would expect no difference in Experiments 1, 3, and 4 (where those frequencies are equal across the two

sources) and a reversed difference for Experiment 2 (where the preferred object is less probable when it is more frequent).

Simulation results for infant crawling patterns are shown in Figure 3, revealing a close match between infant crawling direction and mean samples generated by MCMC operating on the connection weights learned by NPLS. These sample means estimate infants' probability of selecting the favorable vs. unfavorable jar, averaged over 1000 samples for each of 20 networks. In each experiment, the mean selection-probability for the favorable cup is higher than for the un-favorable cup, $p < .0001$.

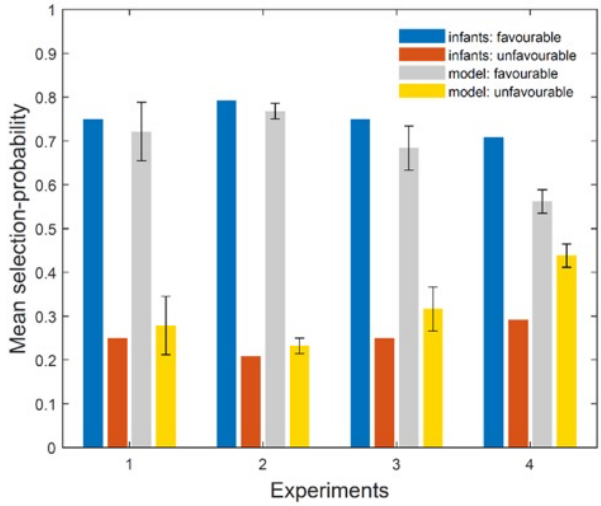

Figure 3. Probability of selecting the favorable (blue) and unfavorable (orange) cups, as demonstrated by infant behavior (Denison & Xu, 2014) and model predictions for probability of choosing the favorable cup (grey) and unfavorable cup (gold), averaged over 20 networks each generating 1000 samples. Error bars denote standard deviations.

Mean MCMC sample values correlate highly with infant search probabilities across the eight conditions of the four experiments, $r(6) = .958$, $p = 1.74\text{E-}4$. Mean MCMC sample values also correlate highly with ground-truth probabilities across the eight conditions of the four experiments, $r(6) = .988$, $p = 4\text{E-}6$. To compute the ground-truth sampling probabilities for Experiments 3 and 4 (corresponding to probability-matching behavior in binary searches), ground-truth probabilities are normalized, by dividing the favorable and unfavorable probabilities by their respective sums.

It is possible that these simple binary probability problems can be learned with only two weights: one from the bias unit, which is always on with an activation value of 1, and another from the input unit. In ongoing work, we devised a mathematical proof of that, identifying unique values of such weights for particular probabilities. We also found that NPLS networks could approximate those ideal weights, even without any hidden units. However, we prefer the model and parameters we present here because it provides an ideal combination of learning accuracy, learning speed, and generality. Our self-organized, recruitment model is more general because, in principle, it can learn any discrete probability distribution with an arbitrary, finite number of outcomes. The more outcome probabilities to learn, the more hidden units are generally recruited.

Across the four simulation experiments, the mean number of hidden units recruited was 2.025, with a range of 0-3, and considerable variation in network topology.

To provide further insight into how and why NPLS networks so readily learn binary probability distributions, we next provide a mathematical analysis. Presented with a frequency ratio N_1:N_0 (with N_1 and N_0 denoting, respectively, the number of preferred and un-preferred objects in a jar), NPLS adjusts its network topology and connection weights so that its output activation x minimizes the sum-of-squared error E, partitioned by presence (target of 1) vs. absence (target of 0):

$$E = \sum_{i=1}^{N_1}(x-1)^2 + \sum_{j=1}^{N_0}(x-0)^2 \qquad (5)$$

Because x^*, the optimal minimizer of E, is given by

$$\frac{d}{dx}E\big]_{x=x^*} = 0 \implies$$

$$2\sum_{i=1}^{N_1}(x^*-1) + 2\sum_{j=1}^{N_0}x^* = 0 \implies \qquad (6)$$

$$x^* = \frac{N_1}{N_1+N_0},$$

the network's output activation comes to closely approximate ground-truth probability. As such, relevant probabilities are mathematically guaranteed to be learned as an emergent property of the system, thus providing a novel explanation of how preverbal infants could learn probabilities without counting and dividing. This proof is generalizable to problems with more than two probabilities.

Discussion

Our model provides a computationally sufficient neural-level explanation of how preverbal infants could learn and use probability distributions without counting and dividing. From the deterministic encoding of the containers of colored objects seen by the infants, the networks directly learn the relevant probabilities without explicitly counting and adding favorable and unfavorable frequencies and without explicitly dividing those two frequencies by total frequencies.

NPLS accomplishes this by autonomously building a network of the appropriate topology and adjusting its connection weights to reduce network prediction error. Other essential features of the model include using an asymmetric sigmoid activation function in the output unit (to keep outputs in the 0-1 range), learning cessation when error reduction stagnates (so that the network does not keep trying to learn probabilistic patterns ad infinitum), and pairing with an MCMC algorithm that uses network weights to probabilistically sample from the learned distribution.

Importantly, the relevant probability distributions are not supplied as learning targets, but rather emerge naturally from neural-network learning of event sequences. Our NPLS model simulates infant visual scanning of the colored-object contents of two jars. Our coding scheme for network learning realistically represents the results of these visual scans by pairing jar identity with relative frequencies of the two colors.

It is likely that other algorithms could also learn the required underlying probability distributions, including the standard method of counting, summing, and dividing of counts by the sum. However, these are skills that infants are known to lack until several years and considerable schooling. Our NPLS model shows the computational sufficiency of techniques that infants could plausibly employ: learning associations between containers and color frequencies by adjusting synaptic weights between neurons and recruiting additional neural units as long as that is helping. And then making inferences in the opposite direction from the preferred color back to the identity of the most probable source jar.

The primary goal of this work is to provide a computationally *sufficient*, neurally plausible account of preverbal infant probabilistic reasoning, demonstrated in Experiments 1-4 from Denison and Xu (2014), without invoking cognitive abilities that are beyond those infants; NPLS achieves that goal.

Our model additionally assumes prior knowledge of physical objects in terms of their solidity, spatial location, and color, consistent with developmental work on core knowledge in infants (Spelke, 2000).

When explaining remarkable infant abilities or any novel results, alternative explanations should be considered. Three alternate ideas can be ruled out because the maximum numerical value they can deal with is only 3-4 items: subitizing (Chi & Klahr, 1975), object files (Feigenson, Carey, & Hauser, 2002), and parallel individuation (Carey, Shusterman, Haward, & Distefano, 2017).

Frequency of preferred items is ruled out by Experiments 1, 2, and 4 (Denison & Xu, 2014). In each experiment, both infant searches and our simulations conformed to probability information and not to information on frequency of their preferred item. The same is true of Experiment 3 (Denison & Xu, 2014) for frequencies of un-preferred items.

Because several other experiments attest that infant understandings of numerosity and ratio are independent of perceptual factors such as area, contour length, and density (McCrink & Wynn, 2007; Wynn, Bloom, & Chiang, 2002; Xu & Spelke, 2000), alternative explanations based on such non-numerical factors are also unlikely here.

Explicit verbal counting could potentially provide accurate magnitude estimations of the large frequencies used in the infant probability experiments (Denison & Xu, 2014), but such counting abilities are still years away from these infants.

There has been some speculation that the Approximate Number System (ANS) could somehow explain the infant results that we simulate here (Denison & Xu, 2014; McCrink & Birdsall, 2015). The ANS is a nonverbal system that allows approximate numerical estimation of collections of items at a glance, yielding magnitude values (Carey et al., 2017; Dehaene, 2009; Feigenson, Dehaene, & Spelke, 2004; Gallistel & Gelman, 1992). Use of the ANS for quantity comparison has been documented in infants as young as 6 months (Feigenson et al., 2004; Xu & Spelke, 2000) and in a range of non-human animals (Agrillo, Piffer, Bisazza, & Butterworth, 2012; Merritt, MacLean, Crawford, & Brannon, 2011). Brain-imaging studies suggest that the ANS engages the intraparietal sulcus of the parietal lobe of human brains (Dehaene, Piazza, Pinel, & Cohen, 2003; Piazza, Pinel, Le Bihan, & Dehaene, 2007).

Perhaps infants could invoke the ANS for approximate magnitude estimation and then apply the standard method of calculating probability to these magnitude estimates: add the two estimates together and divide each frequency estimate by that sum. This seems unlikely because division is a relatively difficult and late-developing skill in children (Gallistel & Gelman, 1992; McCrink & Spelke, 2016). Also, the frequencies used in the unconfounded infant experiments (Denison & Xu, 2014), and in our simulations of those experiments, are considerably larger than the frequencies used in infant experiments on the ANS, which only went up to a maximum of 16 items (Xu & Spelke, 2000). The frequencies used in the infant experiments (Denison & Xu, 2014) and our simulations are far larger, ranging from 4 to 96, with a mean of 30 and standard deviation of 28.

Also noteworthy is the tight accuracy of NPLS probability estimates in matching ground-truth probability computations, and the fact that MCMC sampling from those learned distributions closely matches the infant search patterns (Denison & Xu, 2014). In contrast, the ANS is known to be relatively imprecise, particularly with large numbers, small Weber fractions, and infants (Carey et al., 2017; Feigenson et al., 2004; Xu & Spelke, 2000). Moreover, even if the ANS could provide accurate estimates of these frequencies, it is unclear how the ANS could divide the target frequency by the total frequency of all relevant events, and do so with the precision achieved by the infant search patterns (Denison & Xu, 2014). Moreover, such operations in ANS have not been demonstrated in a working computational model, making it difficult to determine how much explanatory power it has in the area of probabilistic reasoning. These considerations suggest that successful infant probabilistic reasoning is not a product of the ANS.

NPLS is currently the only model that has been demonstrated to produce computationally sufficient simulations of infant learning and use of probability distributions. As such, it is a genuine novelty in theories of early quantitative development. Additional simulations that are currently underway confirm its accuracy with more than two probabilities, including discretized continuous distributions. Several such simulations have also generated novel, testable predictions for infant experiments.

Some other infant experiments measure probabilistic skills with infant surprise at unexpected outcomes instead of search for desired objects (Denison et al., 2013; Teglas et al., 2011;

Xu & Garcia, 2008). In other work, we simulate such differential surprise with the standard measure of network prediction error. Failed prediction triggers more surprise. Although our model can simulate both surprise and search, we are not aware of infant studies using both measures.

In addition, our system is relevant to four other independent lines of research. One of these is the literature on probability matching in a wide variety of animal species, e.g., bees (Greggers & Menzel, 1913), fish (Behrend & Bitterman, 1961), turtles (Kirk & Bitterman, 1965), and apes (De Petrillo & Rosati, 2019; Eckert, Call, Hermes, Herrmann, & Rakoczy, 2018). We will soon be exploring whether experiments with these other species can also be simulated.

There is an active infant literature on transition probabilities, particularly in language learning and vision (Aslin, Saffran, & Newport, 1998; Lany & Gómez, 2008; Saffran, Aslin, Johnson, & Newport, 1999). This is considered to be a distinctly different phenomenon from learning of binary probability distributions (Denison & Xu, 2014), with a distinctly different kind of neural-network modeling (Mareschal & French, 2017).

There is also an emerging literature on statistical summaries (aka ensemble representation (Alvarez, 2011)). Visual cognition is enhanced when human adults quickly summarize the statistical properties (e.g., average and variation) of a collection, affording a precise, compact representation of large collections. This is analogous to our model, in which a probability estimate compactly summarizes numerous presences and absences in event sequences.

As well, there is evidence that adult humans and monkeys can form quick, accurate representations of ratios from visual displays that are coded by neural firing rates in the same brain areas as whole numbers (Jacob, Vallentin, & Nieder, 2012; Kiani & Shadlen, 2009; Matthews & Chesney, 2015). Future research could explore these diverse phenomena, perhaps eventually achieving a unified model.

Our modeling suggests a novel way to begin bridging across Bayesian and neural-network approaches and across different levels of analysis (Marr, 2010). Neural network models operate at a lower, implementational level compared to the higher, computational level of Bayesian models. Each approach often partakes of an intermediate, algorithmic level. Some Bayesian researchers propose bridging across these levels with MCMC sampling (Griffiths, Vul, & Sanborn, 2012). We agree that sampling is important, but also find that accurate constraint-guided sampling is infeasible without some prior learning. The proposed model integrates learning and sampling by reasoning bidirectionally, forward from examples to probability distributions and backwards from probability distributions to examples (Nobandegani & Shultz, 2017, 2018). In hundreds of networks simulating more than a dozen empirical experiments, we have not seen an exception to the idea that successful learning is necessary for accurate sampling. Finally, rather than assuming that probability distributions are innate products of biological evolution, it makes more sense to assume evolution of a powerful learning system that can quickly and accurately register a wide range of novel probability distributions.

Acknowledgements

This research was supported in part by an operating grant to TRS from the Natural Science and Engineering Research Council of Canada.

References

Agrillo, C., Piffer, L., Bisazza, A., & Butterworth, B. (2012). Evidence for two numerical systems that are similar in humans and guppies. *PLoS ONE*, *7*(2).

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8 month old infants. *Psychological Science*, *9*(4), 321–324.

Baluja, S., & Fahlman, S. E. (1994). *Reducing network depth in the cascade-correlation learning architecture*. Carnegie Mellon University. Pittsburgh, PA.

Behrend, E. R., & Bitterman, M. (1961). Probability-matching in the fish. *American Journal of Psychology*, *74*(4), 542–551.

Carey, S., Shusterman, A., Haward, P., & Distefano, R. (2017). Do analog number representations underlie the meanings of young children's verbal numerals? *Cognition*, *168*, 243–255.

Chi, M., & Klahr, D. (1975). Span and rate of apprehension in children and adults. *J Experimental Child Psychology*, *19*, 434–439.

De Petrillo, F., & Rosati, A. G. (2019). Rhesus macaques use probabilities to predict future events. *Evolution and Human Behavior*, *40*(5), 436–446.

Dehaene, S. (2009). Origins of mathematical intuitions: The case of arithmetic. *Annals of the New York Academy of Sciences*, *1156*, 232–259.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3–6), 487–506.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: evidence from 4.5- and 6-month-olds. *Developmental Psychology*, *49*(2), 243–249.

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, *13*(5), 798–803.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335–347.

Eckert, J., Call, J., Hermes, J., Herrmann, E., & Rakoczy, H. (2018). Intuitive statistical inferences in chimpanzees and humans follow Weber's law. *Cognition*, *180*(June), 99–107.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more:

Object files versus analog magnitudes. *Psychological Science*, *13*(2), 150–156.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Gallistel, C. R., & Gelman, R. (1992). Preverbal counting and computation. *Cognition*, *44*, 43–74.

Greggers, U., & Menzel, R. (1913). Memory dynamics and foraging strategies of honeybees. *Behavioral Ecology and Sociobiology*, *32*, 17–29.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: The brain's code for proportions. *Trends in Cognitive Sciences*, *16*(3), 157–166.

Kharratzadeh, M., & Shultz, T. R. (2016). Neural implementation of probabilistic models of cognition. *Cognitive Systems Research*, *40*, 99–113.

Kiani, R., & Shadlen, M. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*, 759–764.

Kirk, K. L., & Bitterman, M. (1965). Probability-learning by the turtle. *Science*, *148*(3676), 1484–1485.

Kolmogorov, A. (1956). *Foundations of the theory of probability* (2d English). New York: Chelsea.

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*(12), 1247–1252.

Mareschal, D., & French, R. (2017). Tracx2: A connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711).

Marr, D. (2010). *Vision: a computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Matthews, P. G., & Chesney, D. L. (2015). Fractions as percepts? Exploring cross-format distance effects for fractional magnitudes. *Cognitive Psychology*, *78*, 28–56.

McCrink, K., & Birdsall, W. (2015). Numerical abilities and arithmetic in infancy. In R. Cohen Kadosh & A. Dowker (Eds.), *The Oxford Handbook of Numerical Cognition* (pp. 258–274). Oxford, England: Oxford University Press.

McCrink, K., & Spelke, E. S. (2016). Non-symbolic division in childhood. *Journal of Experimental Child Psychology*, *142*, 66–82.

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month- old infants. *Psychological Science*, *18*(8), 740–745.

Merritt, D. J., MacLean, E. L., Crawford, J. C., & Brannon, E. M. (2011). Numerical rule-learning in ring-tailed Lemurs (Lemur catta). *Frontiers in Psychology*, *2*(MAR), 1–9.

Moran, P. (1968). *An introduction to probability theory*. Oxford: Clarendon.

Newell, A. (1994). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Nobandegani, A., & Shultz, T. (2017). Converting cascade-correlation neural nets into probabilistic generative models. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1029–1034). Austin, TX: Cognitive Science Society.

Nobandegani, A., & Shultz, T. (2018). Example generation under constraints using cascade correlation neural nets. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 1–6.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: Norton.

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, *53*(2), 293–305.

Saffran, J. R., Aslin, R. N., Johnson, E. K., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Shultz, T. R. (2010). Computational Modeling of Infant Concept Learning: The Developmental Shift from Features to Correlations. In L. M. Oakes, C. H. Cashon, M. Casasola, & D. H. Rakison (Eds.), *Infant Perception and Cognition: Recent Advances, Emerging Theories, and Future Directions* (pp. 125–152). New York: Oxford University Press.

Shultz, T. R. (2017). Constructive artificial neural-network models for cognitive development. In N. Budwig, E. Turiel, & P. D. Zelazo (Eds.), *New Perspectives on Human Development* (pp. 13–26). Cambridge: Cambridge University Press.

Shultz, T. R., & Cohen, L. B. (2004). Modeling age differences in infant category learning. *Infancy*, *5*(2), 153–171.

Shultz, T. R., & Doty, E. (2014). Knowing when to quit on unlearnable problems: another step towards autonomous learning. In J. Mayor & P. Gomez (Ed.), *Computational Models of Cognitive Processes* (pp. 211–221). London: World Scientific.

Spelke, E. S. (2000). Core knowledge. *American Psychologist*, *55*(11), 1233–1243.

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J., & Bonatti, L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059.

Wynn, K., Bloom, P., & Chiang, W. C. (2002). Enumeration of collective entities by 5-month-old infants. *Cognition*, *83*(3).

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in human infants. *Cognition*, *74*, B1–B11.