Query-Based Memory Approximates Rational Induction: Applications to Infant Statistical Learning

Robert Ralston (ralston.123@osu.edu) and **Vladimir Sloutsky (sloutsky.1@osu.edu)**
Department of Psychology, 1835 Neil Avenue
Columbus, OH 43210 USA

Abstract

Query-Based Memory (QBM) models are heavily used in machine learning, though their relevance to human cognition is unclear. In this paper, we explore QBM models through both formal exploration and a simulation study to address this question. We found that QBM models are theoretically motivated, as they approximate rational induction with neurally-plausible mechanisms. Additionally, a simple implementation of the model could readily reproduce four benchmark findings in infant statistical learning. These results provide an encouraging starting point for further research using these formal tools to understand cognition across development.

Keywords: statistical learning; memory; induction; machine learning

Human organisms have a striking ability to adapt to regularities in their environment. This is evident in our ability to solve a variety of statistical problems based on prior experiences. For example, we can categorize objects into disjoint sets, recognize and exploit temporal or spatial patterns, infer the meaning of novel words based on context, and flexibly combine these abilities when it is beneficial to understand the environment. Remarkably, the ability to solve simple versions of these problems emerges early in development, with evidence of category learning by three months (Quinn & Eimas, 1986), statistical learning by two months (Kirkham, Slemmer, & Johnson, 2002; Saffran & Kirkham, 2018), and word learning by six months (Tincoff & Jusczyk, 1999).

In this paper, we draw on recent successes in Natural Language Processing (NLP) to explore a simple but powerful mechanism that could solve many of the problems faced by human learners. Recent advances in NLP have found that transformer models are powerful tools that achieve surprising linguistic performance when trained on large datasets (Devlin, Chang, Lee, & Toutanova, 2019; Brown et al., 2020). These models have, at their heart, the *Attention Layer*, a mechanism integral to achieving state-of-the-art competence in many tasks (Luong, Pham, & Manning, 2015; Vaswani et al., 2017). Despite their name, prior work (Ramsauer et al., 2021; Krotov & Hopfield, 2021; Tyulmankov, Fang, Vadaparty, & Yang, 2021) and our discussion below shows that attention layers are closely analogous to memory retrieval mechanisms. To avoid confusion, we will refer to these layers as Query-Based Memory (QBM) layers throughout the paper.

Below, we first show that QBM Layers implement rational, cluster-based induction (Anderson, 1991). This makes QBM layers a promising candidate for a general learning mechanism, as rational induction has been used as a normative computational model for categorization, statistical learning, word learning, and other cognitive phenomena across development (Perfors, Tenenbaum, Griffiths, & Xu, 2011).

In addition, unlike other implementations of rational models, QBM Layers are differentiable and designed to learn representations gradually via gradient descent. With this learning mechanism, we show that QBM layers can reproduce four benchmark findings in infant statistical learning. While these results do not uniquely support QBM layers as a model of statistical learning, the simulations provide an example of how a rational approach to memory could be implemented to explain human learning in other domains.

Defining Query Based Memory Layers

In the machine learning literature, the most notable QBM mechanism is dot product attention (Luong et al., 2015; Vaswani et al., 2017):

$$QBM(q) = softmax(qK^T)V \tag{1}$$

Note that in high dimensional implementations, an additional constant may be included within the softmax function.

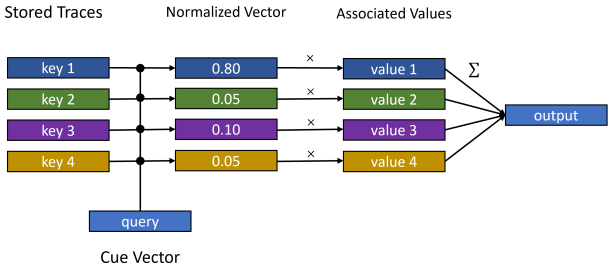

Figure 1: Diagram representing a QBM layer.

This equation has an intuitive interpretation in terms of the retrieval mechanism seen in Figure 1. Under this interpretation, q is a vector of observed features known as a "query". The query is first compared to a number of stored memory traces $K = [k_1, ..., k_k]$ known as "keys" by taking the dot product; $qK^T = [q \bullet k_1, ..., q \bullet k_k]$. The result of this comparison is

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

then put through the softmax function to produce a normalized vector with one entry for each memory trace. Below, we will refer to this as S:

$$S = [s_1, ..., s_k] = [\frac{e^{q \bullet k_1}}{\sum_{i=1}^{k} e^{q \bullet k_i}}, ..., \frac{e^{q \bullet k_k}}{\sum_{i=1}^{k} e^{q \bullet k_i}}] \quad (2)$$

Furthermore, each key is assumed to be associated to a vector, known as a "value," $V = [v_1, ..., v_k]$. The meaning of V can vary depending on context. However, for traditional memory retrieval, values represent features that are associated with the memory traces activated by the cue - i.e., the features to be retrieved from memory. In brief, QBM returns a linear combination of stored values weighted by the softmax of the dot product (S) of the query with each key:

$$QBM(q) = \sum_{i=1}^{k} s_i v_i \quad (3)$$

This expression is reminiscent of memory-based inference models that obtain responses via an exemplar-based or cluster-based mechanism (Kruschke, 1992; Love, Medin, & Gureckis, 2004). In these models, s_i is analogous to the similarity between the current item and a memory trace or cluster-center, while v_i is analogous to the response associated to the i^{th} trace. Furthermore, since S is normalized, this expression recalls the inference mechanism in rational models of induction (Anderson, 1991), where s_i is analogous to the probability that the current item is from the i^{th} element of a partition, and v_i is analogous to the probability of each outcome given an item is from the i^{th} element of the partition. In the sections that follow, we flesh out these analogies, ultimately showing that QBM layers implement a rational model of induction which can be interpreted as memory-based inference.

The Dot Product and Cosine Similarity

To calculate the similarity between a query and the i^{th} stored key, QBM layers make use of the dot product (Eq. 2). At first, the relationship between the dot product and similarity may be opaque, though it has been used by previous models (Hintzman & Ludlam, 1980; Dougherty, Gettys, & Ogden, 1999; Collins, Milliken, & Jamieson, 2020).

However, the dot product is a natural choice to implement a similarity computation because of its relationship to *cosine similarity*, a widely-used similarity measure in psychology and machine learning (Landauer & Dumais, 1997; Günther, Dudschig, & Kaup, 2016). The cosine similarity between two vectors is defined as the cosine of the angle θ between the vectors:

$$sim_{cos}(v_1, v_2) = cos(\theta) \quad (4)$$

Compared to similarity measures that are based on City-Block or Euclidean distance (Nosofsky, 1986), cosine similarity better captures performance in semantic tasks (Günther et al., 2016; Richie & Bhatia, 2021) and is more sensible when vectors become high dimensional (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999).

Note that the dot product and cosine similarity are closely related. If θ_i is the angle between vectors q and k_i:

$$q \bullet k_i = cos(\theta_i) \, \|q\| \, \|k_i\| \quad (5)$$

In the next section, we will show that these two aspects of the dot product (cosine similarity and the scaling constant $\|q\| \, \|k_i\|$) play different conceptual roles in QBM layers.

Inferring Keys from a Query

Consider the vector S constructed in Equation 2. We can rewrite this expression as:

$$s_i = z e^{\|q\| \|k_i\| cos(\theta_i)} \quad (6)$$

where $z = (1/\sum_{i=1}^{k} e^{q \bullet k_i})$ is the constant from Equation 2.

This formulation is highly suggestive. Recall that s_i is the outcome of comparing a query with the ith stored memory trace. If we consider s as a function of θ_i, then $s(\theta)$ is proportional to a centered *von Mises density* (Damien & Walker, 1999), a receptive field-like distribution defined over circular quantities such as the angle between vectors:

$$p(\theta|\kappa) = \frac{e^{\kappa cos(\theta)}}{2\pi I_0(\kappa)} \quad (7)$$

Here, κ is a positive dispersion parameter (analogous to $\frac{1}{\sigma}$ in a normal distribution), $\theta \in [0, 2\pi)$, and $I_0(\kappa)$ is a modified Bessel function of the first kind of order 0. See Figure 2 for a visualization of this distribution for various values of κ.

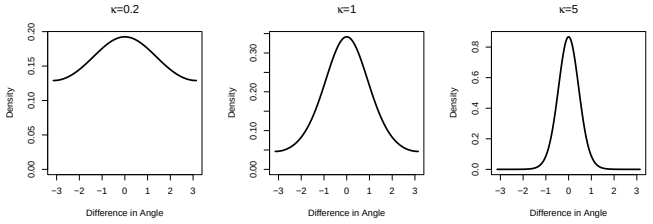

Figure 2: Centered von Mises distributions.

This distribution is closely analogous to the expression for S in Equation 6. When the dispersion parameter

$$\kappa_i = \|q\| \, \|k_i\| \, , \quad (8)$$

then

$$s_i = 2\pi z I_0(\kappa_i) p(\theta_i|\kappa_i) \quad (9)$$

Note that, here and in what follows, we exclude the case where the length of a vector is precisely equal to zero.

The ability to express elements of S in terms of a probability density is not a coincidence. In fact, we can show that the vector S is secretly implementing Bayes' Theorem to infer $Pr(k_i|q)$, the probability that a distribution associated to the ith key vector would produce the query.

The Probabilistic Interpretation

Assume that an individual has stored accurate memory traces corresponding to a discrete collection of entities (e.g., objects). They are then shown one of the previously-seen entities and must determine which of the entities it is.

To solve this problem, they may reason that their current observations are produced by one of the previously stored items according to a noisy process. To formalize this, they could represent the memory traces generated by previously-seen entities as a probability distribution centered on the stored traces. Then they could calculate the probability that each of the entities was producing their observations.

The same logic motivates the probabilistic interpretation of QBM layers with some technical caveats. First, a rational QBM learner needs to specify what kind of noise is affecting their current observations. This noise model could be arbitrarily complicated and dependent on prior knowledge about each stimulus dimension. Alternatively, a learner could take a much simpler approach. Whether due to ignorance or a desire for parsimony, they may reason that the direction that the vector representing their current observations (the query) differs from that representing a stored trace (a key) does not matter; the likelihood of a query only depends on the distance between the query and the key. This simplifying assumption implies that $p(q|k)$ is radially symmetric. Since $p(q|k)$ is radially symmetric, the QBM learner can model this probability with a unidimensional distribution $p(q|k) = p(distance(q,k)|mean = 0,\ dispersion = \kappa)$.

Second, since items in memory have the potential to be very high dimensional, cosine similarity provides a sensible method to compare the current item (the query) with those stored in memory (keys). As discussed above, cosine similarity is closely related to angular distance. Thus, the desired probability distribution corresponding to each key vector would be a probability distribution over possible angular distances that could occur between query and key vectors.

Since an angle of 0 radians is identical to an angle of 2π radians, angular distance is circular. One receptive field-shaped distribution over circular dimensions is the von Mises distribution (Eq. 7), which closely approximates the wrapped Gaussian distribution. This provides a rational basis to use von Mises distributions to model $p(q|k)$, the likelihood of one's current observations given that they are observing item k. Therefore, in what follows:

$$p(q|k_i) = p(\theta_i|\kappa_i) = \frac{e^{\kappa_i cos(\theta_i)}}{2\pi I_0(\kappa_i)} \tag{10}$$

Since this expression gives us the likelihood of our current observation (query) given we are observing each item (key), we can infer $Pr(k|q)$, the rational confidence that we are observing each entity, using Bayes' Theorem:

$$S = [Pr(k_1|q), ..., Pr(k_k|q)] \tag{11}$$

$$= [\frac{Pr(k_1)p(\theta_1|\kappa_1)}{\sum_j Pr(k_j)p(\theta_j|\kappa_j)}, ..., \frac{Pr(k_k)p(\theta_k|\kappa_k)}{\sum_j Pr(k_j)p(\theta_j|\kappa_j)}] \tag{12}$$

We can show that QBM layers implement this computation with a specific choice of the prior distribution for each key. In Equation 9, we showed that s_i is expressed in terms of $p(\theta_j|\kappa_j)$. Therefore, if there exists a valid probability distribution $Pr(k_i)$ such that

$$2\pi z I_0(\kappa) = \frac{Pr(k_i)}{\sum_j Pr(k_j)p(q|k_j)}, \tag{13}$$

then QBM layers implement the probabilistic inference described above.

Theorem. *Equation 13 holds if* $Pr(k_i) = \frac{I_0(\kappa_i)}{\sum_j I_0(\kappa_j)}$.

Proof. This follows straightforwardly from the left side of Equation 13.

$$2\pi z I_0(\kappa_i) = \frac{2\pi I_0(\kappa_i)}{\sum_j e^{\kappa_j cos(\theta_j)}} \tag{14}$$

$$= \frac{2\pi I_0(\kappa_i)(1/\sum_k I_0(\kappa_k))}{\sum_j [e^{\kappa_j cos(\theta_j)}(I_0(\kappa_j)/I_0(\kappa_j)](1/\sum_k I_0(\kappa_k))} \tag{15}$$

$$= \frac{Pr(k_i)}{\sum_j \frac{e^{\kappa_j cos(\theta_j)}}{2\pi I_0(\kappa_j)} Pr(k_j)} \tag{16}$$

$$= \frac{Pr(k_i)}{\sum_j Pr(k_j)p(q|k_j)} \tag{17}$$

□

Therefore, we have shown that $softmax(qK^T)$ from Equation 1 implements a rational process, determining the probability that an observed cue (query) is generated by the entity corresponding to each stored memory trace (key).

Memory-Based Induction

Above, we showed that the normalized vectors computed in the first half of Equation 1 implement a probabilistic model, where the result is

$$S = [Pr(k_1|q), ..., Pr(k_k|q)]. \tag{18}$$

Recall that each stored trace (k_i) is also associated with a value (v_i), and the QBM layer ultimately returns a linear combination of values weighted by components of S (Equation 3). In this section, we will show that this is equivalent to an approximately rational model of induction (Anderson, 1991).

To investigate a rational approximation to induction, Anderson (1991) gave a cluster-based account which has been substantially expanded (Sanborn, Griffiths, & Navarro, 2006) while retaining the general characteristics described below. This approach claims that organisms first categorize entities into disjoint clusters k according to their observed features q. Then, to perform induction, organisms determine the probability of an unobserved value y according to the expression

$$Pr(y|q) = \sum_i Pr(k_i|q)Pr(y|k_i). \tag{19}$$

Thus, on this view, induction is a two step process. First, organisms determine the probability that each cluster generated the observed features. Then, they combine this information with the probability that each cluster will produce the target.

According to our above discussion, $s_i = p(k_i|q)$. With this equality, Anderson's equation (19) is strikingly similar to Equation 3 above. Specifically, if we interpret the value associated with the ith stored trace (v_i) as the probability distribution $Pr(y|k_i)$, then QBM layers can be said to implement Anderson's approximation to rational induction.

This is sensible in the memory context described above. If each v_i is a one-hot vector (a vector of zeros with a 1 in the i^{th} place), then V will be an identity matrix. In this case, $QBM(q)$ is the distribution over keys shown in Equation 18. Thus, the output of a QBM layer would be a posterior distribution over all of the entities which could be producing one's current observations - i.e., a rational retrieval mechanism.

Furthermore, this mechanism can be extended well beyond memory retrieval. To show this, the simulations below use a QBM mechanism to learn sequential dependencies, such as those from infant statistical learning. In this context, query vectors represent an infant's current observations, while each key corresponds to a potential next item which may be observed. Through exposure to a sequence of stimuli, infants adjust the position of query and key vectors, such that the query produced by item i is near key j if item j is likely to follow item i (this process is described in the next section). Thus, s_j will be large when item j is expected to occur next. As in the previous paragraph, if each v_i is one-hot, then Equation 1 will produce a posterior distribution over items which may occur next - i.e., a prediction about the future.

We believe that this implementation is sensible in the statistical learning context. However, note that in other contexts, including NLP, values may not be one-hot or normalized, leading to cases where values do not represent valid probability distributions. Though beyond the scope of the current paper, we note that Equation 3 finds the expectation of v, allowing the model to infer values in a sensible way even when this constraint is violated. However, in these cases, $QBM(q)$ do not necessarily represent a valid posterior distribution.

Learning with Query-Based Memory

In the previous sections, we argued that QBM layers implement rational induction and can be applied to problems such as memory retrieval and prediction. In this vein, other rational models of induction have been proposed (Sanborn et al., 2006; Lloyd, Sanborn, Leslie, & Lewandowsky, 2019). One challenge for these models is specifying a psychologically-plausible learning mechanism which adjusts model parameters according to experience.

One of the strengths of Query-Based Memory models is their ability to learn via gradient descent. In NLP, this is accomplished by learning embeddings of query, key, and value vectors (as well as by adjusting the weights of additional (dense) layers) (Vaswani et al., 2017). For example, if model error would be reduced by having cue vector q_i elicit a response more like associated value v_j, the model could move the embedded q_i in the direction of the embedded k_j in order to increase s_j and, thus, more strongly activate v_j. As outlined above, this is particularly useful in the statistical learning context, as the query corresponding to item i can be moved nearer to the j^{th} key when item j appears after item i.

In the most general case, where embeddings of cues, traces, and associated values are learned independently, this is accomplished by learning embedding matrices W_q, W_k, and W_v, via gradient descent, where:

$$QBM(q) = softmax((qW_q)(KW_k)^T)VW_v \tag{20}$$

However, in many applications, one or more of these matrices may be identified or omitted. For example, in our simulations, we do not use the value embedding, W_v. This omission paired with the requirement that V is an identity matrix corresponds to the case where activation of the i^{th} key indicates that item i likely occurs next. This also allows each value to always represent a valid probability distribution ($\|v_i\| = 1$ for all i), fully implementing Equation 19.

Additionally, drawing on the analogy between QBM layers and memory models, QBM layers can also learn by storing a new key and its associated value (i.e., adding a row to K and V). In the simulations that follow, we focus on learning embeddings rather than storing new traces due to infants' generally poor episodic memory (Newcombe, Lloyd, & Ratliff, 2007). However, this is an exciting avenue for further research comparing QBM models to other memory models as well as data from older age groups.

Simulations of Infant Statistical Learning

In the previous sections, we introduced query-based memory layers and showed that these layers implement rational induction. In addition, unlike other models of rational induction, QBM layers can be gradually trained with gradient descent in order to learn item embeddings that allow for inferences beyond memory retrieval. We now turn to the specific case of infant statistical learning to show that QBM layers readily produce patterns observed in experimental studies.

We chose the case of infant statistical learning for two reasons. First, as presented above, QBM layers do not possess a goal-directed attention mechanism. Since selective attention is an important aspect of learning for older children and adults (Deng & Sloutsky, 2016), QBM layers would need to be augmented with additional mechanisms to explain these findings. Second, we wanted to test the ability of the model to explain learning without the storage of new episodic traces. Since infants have generally poor episodic memory, this was a good test-case for the efficacy of this mechanism.

Prior research has attempted to model statistical learning in a number of ways, including chunking and assessing the familiarity of items in memory (Perruchet & Vinter, 1998; French, Addyman, & Mareschal, 2011), recurrent neural network architectures (Cleeremans & McClelland, 1991), and

others. The simulations below do not uniquely support QBM layers, as many models can reproduce these results. Instead, the goal of the current simulations is to show that a mechanism designed to solve a seemingly different problem (memory retrieval) can solve a much larger class of problems (prediction). Given that the hippocampus, a brain region associated with memory storage and retrieval (McClelland, McNaughton, & O'Reilly, 1995), has recently been implicated in predictive processing and statistical learning (Schapiro, Kustner, & Turk-Browne, 2012; Schapiro, Turk-Browne, Norman, & Botvinick, 2016; Schapiro, Turk-Browne, Botvinick, & Norman, 2017), we believe that this is an important insight.

Model Implementation

All of the statistical learning phenomena considered here involve sequences of discrete sets of stimuli, such as sequences of phonemes. For explication, assume participants are shown a sequence of items, $x_1, x_2, \ldots$. During each stimulus presentation, we assume that infants try to predict the next stimulus that will occur. Thus, the observation of item x_t would cause an infant to generate a probability distribution over items representing their prediction of what will occur at x_{t+1}.

As we described above, this was implemented in a QBM layer by using a vector representing the current stimulus as a query, and using a one-hot representation of the i^{th} item as the i^{th} value. Additionally, since the model learned an embedding for queries (W_q) and keys (W_k), it was sensible to use one-hot vectors for queries (x_t below) and keys (K) as well.

With these stipulations, a QBM layer could implement rational induction:

$$Pr(x_{t+1}|x_t) = softmax((x_t W_q)(KW_k)^T)V \qquad (21)$$

This mechanism is presented schematically in Figure 3. At the start of learning, the embedding matricies W_q and W_k were randomly initialized. Then, gradient descent was used at each time step to change the embeddings, ultimately reducing the prediction error. Note that, for each simulation, the model was exposed to the same number of stimuli as infants in the original studies.

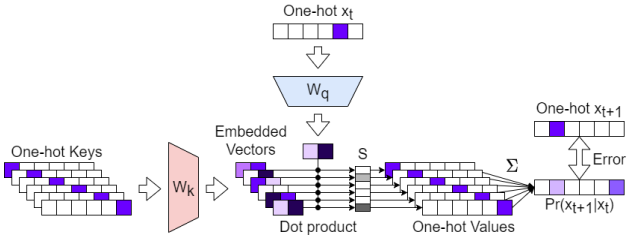

Figure 3: A QBM layer applied to statistical learning.

The model was tested by having it predict sequences of generalization items. This followed a typical habituation paradigm, where infant learning is inferred by differential looking times to novel sequences of stimuli after habituation (Colombo & Mitchell, 2009). For our purposes, we inferred that the model was infant-like if it a) produced differences in prediction error for novel sequences in the same conditions that infants show significant differences in looking time, and b) failed to produce differences in prediction error where infants fail to produce differences in looking time.

Finally, note that, in the simulations below, we do not fit the model. In this implementation, the model only has two free (hyper)parameters, the learning rate (set at 0.01), and the number of dimensions in the embedding (set at 2). With this limited setup, we hope to show that the model produces infant-like learning with few free parameters.

Methods and Results

All simulations were implemented in Python using Tensorflow (Abadi et al., 2015). Embedding weights were initialized using the Glorot Uniform initializer algorithm (Glorot & Bengio, 2010). In addition, we chose to use a 2 dimensional item embedding across all experiments (i.e., our embedding matrices had 2 columns) in order to simulate infants' limited representational capacity (this forced items to interfere with each other even under optimal performance). Finally, to simulate the effect of different initial embeddings, all simulations were run 100 times and model results were collectively considered as a sample for statistical inference.

During learning, stimulus presentation to the model was matched to the original papers, and W_q and W_k were altered via stochastic gradient descent. Because the model outputs a probability distribution, we used categorical cross-entropy loss as an error signal, comparing model predictions to a one-hot vector representing the true next stimulus. During testing, no updates were performed. To assess model surprise during test conditions, we calculated the mean loss per sample for each condition reported in the original paper, and considered this value as a measure of surprise. We then used paired sample t-tests to assess the discrimination of the model.

Word Segmentation: Saffran, Aslin, and Newport (1996) In this study, 8-month-old infants were presented speech streams of syllables containing statistical regularities. During a familiarization phase, infants heard speech consisting of many instances of four novel words made from combinations of 3 syllables. In Experiment 1, infants were able to discriminate between previously-heard words and novel words constructed using the same syllables, $t(23) = 2.3, P < 0.04$. Then, in Experiment 2, infants were able to discriminate previously heard words from repetitions of "part words" $t(23) = 2.4, P < 0.03$. Part words were generated using the final syllable from one word and the first two syllables of another word.

The model was trained on sequences equal in length to those used for infants. Paired samples t-tests revealed that the model was more surprised by novel words than previously-heard words in Experiment 1, $t(99) = 18.40, P < 0.001$, and in Experiment 2, $t(99) = 16.81, P < 0.001$. Therefore, the model showed the same pattern of discrimination as infants.

Artificial Grammar: Gomez and Gerken (1999) This study presented 1-year-old infants with speech generated using an artificial grammar with five words. During an acquisition phase, participants were exposed to sentences generated from the grammar. Then, during test, participants were presented grammatical and ungrammatical strings. Experiment 1, simulated here, tested whether infants could discriminate grammatical strings from strings with an invalid endpoint. A repeated measures ANOVA found that infant looking times were affected by grammaticality, $F(1,15) = 9.09, p = .01$.

To simulate the experiment, strings were created by including all valid words as well as a start point and endpoint symbol, appearing at the beggining and end of each sentence respectively. A paired samples t-test found that the model was more surprised by strings with an invalid endpoint than by novel grammatical strings, $t(99) = 6.83, P < .001$.

Nonadjacent Dependencies: Gomez and Maye (2005) In this study, 12 month old infants were presented with strings of novel words that contained predictable nonadjacent dependencies - predictable regularities that span longer than a single transition. They were then presented with novel utterances that obeyed or violated the nonadjacent dependencies. We simulated results from Experiment 2B, where infants did not show evidence of discrimination between novel grammatical and ungrammatical sentences, $t(23) = 0.52, P = .950$.

As in the previous simulation, we included symbols denoting the start and end of a sentence. Otherwise, model exposure was identical to stimuli shown to participants. The model also failed to detect a difference between grammatical and ungrammatical strings, $t(99) = 1.09, P = .277$.

The failure of both infants and the model to learn nonadjacent dependencies is revealing. The model fails to learn because, in this simple implementation, there is no mechanism to incorporate information from more than one time step prior. This could also be the case for young infants, who often fail in tasks that require retaining information through time (Ross-Sheehy, Oakes, & Luck, 2003).

Combined Results In sum, Query-Based Memory was able to reproduce four findings from infant statistical learning. The model was able to learn transition probabilities during word segmentation as well as discriminate grammatical and ungrammatical utterances in an artificial grammar learning task. However, it was not able to learn every pattern, and struggled to learn nonadjacent dependencies in a task where 12 month old infants show little evidence of discrimination.

Overall, these findings provide an encouraging sign that Query-Based Memory layers reproduce some important phenomena in statistical learning via representational change. We do not claim that these results are unique to QBM models. However, they give a solid starting point to continue investigating this mechanism and other capabilities of QBM models.

Discussion

In this paper, we showed that Query-Based Memory (QBM) models, widely used to solve machine learning problems in NLP, are also rational models of induction. We proved that QBM layers are equivalent to a rational probabilistic model combined with a cluster-based inference mechanism. However, unlike other approximately rational models, QBM models are differentiable, and can be used to learn incrementally via gradient descent. In particular, this allows QBM layers to learn representations of items that are amenable to inference using the specified probabilistic model.

Formally, these findings have several consequences. First, though other authors have interpreted the outcome of an attention layer as a probability distribution (Vinyals, Blundell, Lillicrap, Wierstra, & Kavukcuoglu, 2016), we are the first to derive the representation of dot product attention as a radial von Mises mixture model. This analogy may prove to be helpful for machine learning applications, as it allows us to ask whether typical implementations of QBM layers may benefit from reparameterizing the model to allow for better performance.

In addition to our formal findings, we also showed that QBM layers are able to reproduce four important findings in infant statistical learning. This suggests that infants may succeed in these tasks without storing new memory traces. Instead, our model was able to learn by altering the pattern of activation elicited by items, as well as the pattern of activation needed to retrieve an association, via self-supervised learning. Modifying item representations is a important aspect of perceptual learning (Goldstone, 1998) and has been hypothesized to occur in adults as a product of semantic exposure (Hofmann, Muller, Rolke, Radach, & Biemann, 2020) and category learning (Goldstone, Lippa, & Shiffrin, 2001). Future work can attempt to correlate the representational changes predicted by QBM models with human behaviors.

Additionally, QBM Layers have been shown to be equivalent to modern Hopfield Networks (Ramsauer et al., 2021), which have been suggested as a model of hippocampal function, specifically for pattern completion in subfield CA3 (Rolls, 2013; Janarthanam, Vishwanath, & Shanthi, 2020). Recent findings suggest that the hippocampus plays a role in statistical learning (Schapiro et al., 2016), with CA3 involved in the prediction of upcoming items (Schapiro et al., 2012). Future work can explore the relationship between QBM layers and hippocampal function as well as compare QBM layers to other explanations of hippocampal development and its relation to statistical learning (Schapiro et al., 2017).

Finally, though our simulation study focused on learning via representational change, QBM models can also learn by storing new exemplars. In the future, we hope to expand the simple QBM model to include other cognitive processes such as memory storage and selective attention. Including these additional mechanisms will allow QBM models to be applied to more complex tasks, as well as tasks involving older age groups where these mechanisms play a crucial role.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

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

Beyer, K. S., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In *Icdt.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *ArXiv*, *abs/2005.14165*.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235-253.

Collins, R. N., Milliken, B., & Jamieson, R. K. (2020). Minerva-de: An instance model of the deficient processing theory. *Journal of Memory and Language*, *115*, 104-151.

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, *92*, 225-234.

Damien, P., & Walker, S. G. (1999). A full bayesian analysis of circular data using the von mises distribution. *Canadian Journal of Statistics-revue Canadienne De Statistique*, *27*, 291-298.

Deng, W., & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, *91*, 24-62.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, *abs/1810.04805*.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). Minerva-dm: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180-209.

French, R. M., Addyman, C., & Mareschal, D. (2011). Tracx: a recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614-636.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the international conference on artificial intelligence and statistics.*

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585-612.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*, 27-43.

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109-135.

Gómez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*, 183-206.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*, 626-653.

Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory & Cognition*, *8*, 378-382.

Hofmann, M. J., Muller, L., Rolke, A., Radach, R. R., & Biemann, C. (2020). Individual corpora predict fast memory retrieval during reading. *ArXiv*, *abs/2010.10176*.

Janarthanam, V. V., Vishwanath, S., & Shanthi, A. P. (2020). A biologically plausible network model for pattern storage and recall inspired by dentate gyrus. *Neural Computing and Applications*, *32*, 13289–13299.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*, B35-B42.

Krotov, D., & Hopfield, J. J. (2021). Large associative memory problem in neurobiology and machine learning. *ArXiv*, *abs/2008.06996*.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lloyd, K., Sanborn, A. N., Leslie, D., & Lewandowsky, S. (2019). Why higher working memory capacity may help you learn: Sampling, search, and degrees of approximation. *Cognitive Science*, *43*, e12805.

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

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *ArXiv, 1508.04025*.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419-457.

Newcombe, N. S., Lloyd, M., & Ratliff, K. (2007). Development of episodic and autobiographical memory: a cognitive neuroscience perspective. *Advances in Child Development and Behavior*, *35*, 37-85.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-61.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to bayesian models of cognitive development. *Cognition*, *120*, 302-321.

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of Memory and Language*, *39*, 246-263.

Quinn, P. C., & Eimas, P. D. (1986). On categorization in

early infancy. *Merrill-Palmer Quarterly*, *32*, 331-363.
Ramsauer, H., Schafl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., ... Hochreiter, S. (2021). Hopfield networks is all you need. *ArXiv*, *abs/2008.02217*.
Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*, e13030.
Rolls, E. T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, *7*.
Ross-Sheehy, S., Oakes, L. M., & Luck, S. J. (2003). The development of visual short-term memory capacity in infants. *Child Development*, *74*, 1807-1822.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.
Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181-203.
Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the cognitive science society.*
Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*, 1622-1627.
Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*.
Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*, 3-8.
Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*, 172-175.
Tyulmankov, D., Fang, C., Vadaparty, A., & Yang, G. R. (2021). Biological learning in key-value memory networks. *ArXiv*, *abs/2110.13976*.
Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *ArXiv*, *abs/1706.03762*.
Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., & Kavukcuoglu, K. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, *29*.