UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Determinantal Point Processes for Memory and Structured Inference

Permalink

https://escholarship.org/uc/item/0rx7s5dp

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 42(0)

Authors

Frankland, Steven M.
Cohen, Jonathan D.

Publication Date

2020

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Determinantal Point Processes for Memory and Structured Inference

Steven M. Frankland
steven.frankland@princeton.edu
Princeton Neuroscience Institute
Princeton, NJ

Jonathan D. Cohen
Princeton Neuroscience Institute
Princeton, NJ

Abstract

Determinantal Point Processes (DPPs) are probabilistic models of repulsion, capturing negative dependencies between states. Here, we show that a DPP in representation-space predicts inferential biases toward *mutual exclusivity* commonly observed in word learning (mutual exclusivity bias) and reasoning (disjunctive syllogism) tasks. It does so without requiring explicit rule representations, without supervision, and without explicit knowledge transfer. The DPP attempts to maximize the total "volume" spanned by the set of inferred code-vectors. In a representational system in which combinatorial codes are constructed by re-using components, a DPP will naturally favor the combination of previously un-used components. We suggest that this bias toward the selection of volume-maximizing combinations may exist to promote the efficient retrieval of individuals from memory. In support of this, we show the same algorithm implements efficient "hashing", minimizing collisions between key/value pairs without expanding the required storage space. We suggest that the mechanisms that promote efficient memory search may also underlie cognitive biases in structured inference.

Keywords: mutual exclusivity; determinantal point process; memory; binding; compositionality; probabilistic models

Background and Motivation

Imagine that Tim and Tom are playing an Argentinian board game called El Estanciero. Tim won. What happened to Tom? Although you may not be certain, you don't need any more information to infer that it's likely that Tom lost. This is true despite the fact that you have no familiarity with the particulars of the game or the people involved. This reflects an inferential bias toward *mutual exclusivity* (ME): the tendency to map individuals to $1/n$ possible relational positions, and not to multiple positions in the same instance. Although, in the El Estanciero example, you came equipped with rich knowledge about the relational structure of *games* that you could import, ME inferences have been observed surprisingly early in human development (Halberda 2003, Cesana-Arlotti et al. 2018), and in non-human species (Pepperberg et al. 2019), and aren't constrained to binary relations.

For example, a classic finding in developmental psychology is that, all else being equal, young children prefer to map a novel word ("zurp") to a novel referent (cathode-ray tube), rather than mapping many words to the same object, or many objects to the same word (Markman & Wachtel, 1988; Halberda, 2003), a phenomenon known as the *mutual exclusivity bias* (ME) in word-learning. At the same time, work in a different cognitive domain has found that pre-verbal infants assume that an individual object can not be in two places at the same time, and thus make inferences that resemble a formal disjunctive syllogism (*A or B (but not both). Not A. Therefore, B*) (Cesana-Arlotti et al. 2018; See also Mody & Carey, 2016)).

Here, we argue that the class of ME inferences can be fruitfully modeled using a Determinantal Point Process (DPP) operating over a representational space. DPPs are probabilistic models of repulsion between states: the more similar two states are, the less likely they are to co-occur (See Figure 1). DPPs originated in statistical physics to model the location of fermions at thermal equilibrium (Macchi, 1975), but have since been extended to other branches of mathematics and machine learning (Kulesza & Taskar, 2012). In machine learning, they have recently gained traction in the generation of sets of samples when sample *diversity* is desirable, such as recommender systems looking to present a broad sample of item-types to users ((Kulesza & Taskar, 2012, Gillenwater et al. 2012).

Here, we consider the inferential biases afforded by DPPs over a representational space. Specifically, we show (1) that when representations of possible combinations (e.g., word/object or location/object combinations) are combinations of re-usable codes, DPPs naturally predict the mutual exclusivity bias in word learning and reasoning by disjunctive syllogism. We then (2) suggest that these inferential biases may owe to basic desiderata placed on the data structures employed by an efficient memory system, and provide evidence that DPPs effectively navigate a space/time tradeoff regarding storage space and access time in encoding and retrieval.

General Methods

We focus on two well-studied cases of mutual exclusivity in cognitive science: The "mutual exclusivity bias" in word learning (Markman Wachtel, 1988; Merriman Bowman,

©2020 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

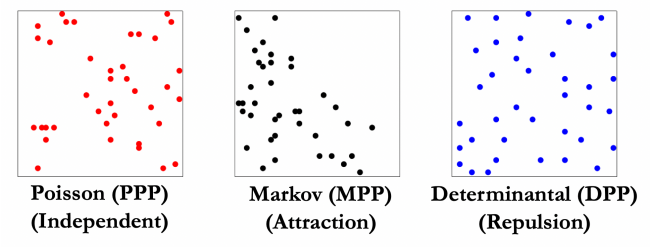

Figure 1: Sampling in a plane for three types of point processes. PPP's contain no relational information, and therefore, although likely to be spread across the space, can be susceptible to "clumping". MPPs, by contrast, carry relational information about the similarity between states, directly promoting such clumping. DPPs use the same similarity representation as MPPs, but, critically, *repel* similar states, ensuring that samples are well distributed across the space.

1991; Halberda, 2003) and the ability to complete disjunctive syllogisms (Mody & Carey, 2016; Cesana-Arlotti et al. 2018). We first describe the modeling approach in abstract terms, explaining the features that are common to both use-cases.

DPP framework. Our model assumes a point process P operating over a finite ground set Y of discrete items. Here, the items are possible *representations* (i.e., encodings) and more specifically, representations of particular combinations (e.g., word/object or location/object combinations). We will often use the generic terms "keys" and "values" when talking about the representational components, motivated largely by the connections we wish to make to memory.

The process defines a probability of selecting particular subsets (S) of these combinations drawn from the ground set.

$$P(S \subseteq Y) = det(K_S)$$

where K is an NxN positive semi-definite kernel matrix whose entries encode pairwise similarities between N possible discrete states, where $N = m$ keys X n values. For all analyses reported here, we use a linear kernel, computed as a normalized inner product of each combination of the mXn key/value vectors under consideration (See Figure 2 for illustration).

DPPs select a particular configuration of items so as to maximize the determinant (det) of the corresponding sub-matrix of K, indexed by S. (Macchi, 1975; Kulesza Taskar, 2012). Thus, the central computation in the current case is $arg\ max_S\ det(K_S)$.

Geometrically, we can think of this determinant as the *volume* spanned by the parallelpiped of the code vectors. The more similar a set of vectors (small determinant), the less likely they are to co-occur in a set. The less similar the vectors (larger determinant), the more likely they are to co-occur in S. This enables the modeling of repulsion between possible states; it is central to the current work and its ability to model aspects of higher-level cognition.

Here, the relevant representational states are concatenations of two types of code vectors corresponding to pre-factorized representations. As we noted above, these factors may be thought of as "keys" and "values", or alternatively "relations" and "content". Concretely, however, in the situations we explore, they are representations of *words* and their *referents* in the ME-bias case, and *spatial locations* and *objects* in the disjunctive syllogism case. To generate the code for a possible combination, we simply concatenate vectors for the key/value components, keeping the ordering and codes consistent across uses, consistent with compositionality.[1] Although we report simple key/value concatenations here, we obtain the same results both by summing key/value representations.

Although finding the maximum a posteriori (MAP) subset in a DPP is NP-hard (Kulesza & Taskar, 2012), there exist greedy methods that can effectively approximate it (Gillenwater et al. 2012, Han et al. 2017). However, here, we make the simplifying assumption that Y is itself a subset of the vastly larger set of possible items that could have been under consideration. For present purposes, we assume that this context-dependent restriction of the possibility space can be carried out by standard attentional mechanisms. Within this small-cardinality space, we are able to exhaustively search for the MAP subset (the sub-matrix that maximizes the determinant). However, generating more plausible heuristic methods that can scale to larger spaces remains a focus for future work.

Throughout, we compare the DPP to two alternative point process models in order to emphasize the conceptual contribution of the DPP. First, a Poisson Point Process (PPP), which assumes no similarity kernel K, treating each item as independent. Here, we use $P(S \subseteq Y) = \prod_{i \in y} p_i \prod_{i \notin y} (1 - p_i)$, where p is a flat prior across states. This is random uniform selection. Second, a generic Markov Point Process (MPP), which selects items based on the kernel K used for the DPP, but selecting directly on the similarity scores of the sub-matrix, rather than its determinant. In this, the MPP over K can be considered in opposition to the DPP, favoring items that are nearby in the code-space rather than far apart. Taken together, one can think of these three models as capturing the possibility of (a) random inference (PPP), (b) inference by similarity (MPP) and (c) inference by repulsion (DPP). See Figure 1. We note that we are not choosing to compare the

[1] We note that, although this encoding framework is simple, it is motivated by empirical evidence concerning the nature and organization of the projections from the mammalian entorhinal cortex (EC) to the hippocampal sub-fields: a key circuit both for simple forms of reasoning and memory (Zeithamova et al., 2012). Specifically, a medial region of EC contains low-dimensional representations of the spatial structure of the environment, while a lateral region encodes sensory content (Behrens et al. 2018). These separate representations are believed to then be bound together in the hippocampus in order to encode different structure/content combinations (Whittington et al. 2018).

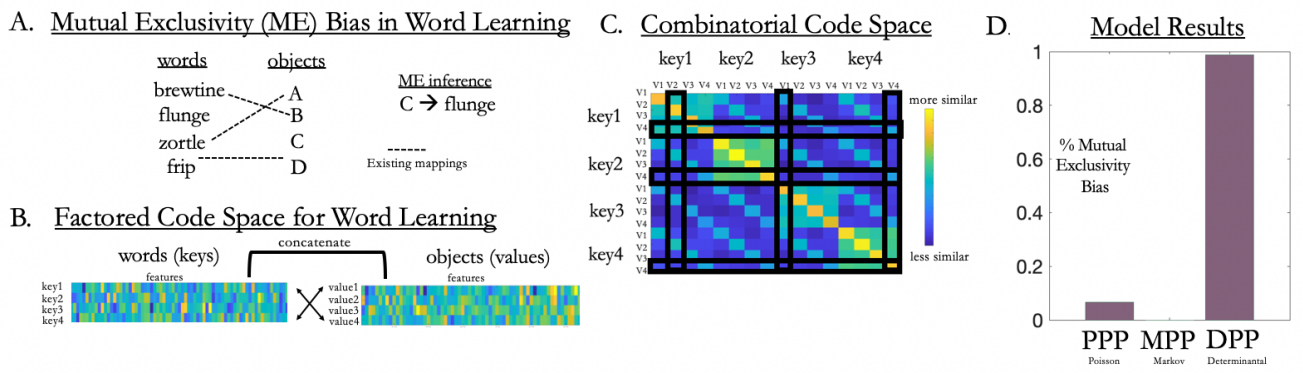

Figure 2: **(A)**. Example of the word learning problem. The model's task is to select a word/object mapping, conditioned on 3 existing associations. Humans tend to map the held-out word to the held-out object. **(B)** shows the representations used by the models to guide inference. We assume separate, but combinable, codes for words (keys) and objects (values). The square matrix in **(C)** represents pairwise similarities between possible word/object combinations. Brighter colors reflect more similar combinatorial codes, darker colors less similar codes. Black bars across rows and columns reflect a hypothetical subset of word/object mappings, as in A. **(D)** We evaluate the probability that the held-out word is mapped to the held-out object (mutual exclusivity bias) across 1000 simulations with different word/object codes. A DPP naturally selects the novel unused word and un-used object, exhibiting the mutual exclusivity bias.

DPP to these other point process models because we believe that MPPs and PPPs are *a priori* particularly plausible models of the relevant types of structured inference. Instead, we believe they are useful to highlight fundamental properties of the DPP (e.g., repulsion). We believe that first comparing the predictions of a DPP against these clearly situates the DPP in a consistent conceptual framework for expository purposes.

Mutual Exclusivity Bias in Word Learning. The "mutual exclusivity (ME) bias" in word learning refers to the empirical observation that, all else being equal, young children and adults prefer to map a novel word to a novel referent. They prefer this both to mapping many words to the same object, or many objects to the same word (Markman & Wachtel, (1988); Merriman & Bowman, (1989); Halberda, (2003); Lake, Linzen, Baroni (2019)). Here, we suggest that this inferential bias follows directly from the consideration of an associative encoding system that selects which codes to bind under a Determinantal Point Process. A learner that performs inference using a DPP will prefer combinations that maximize the total volume of the representational space. Under the assumption that components re-use codes across possible uses (consistent with compositionality), a DPP naturally favors combinations of previously *un-used* words and objects.

We model a case involving 4 words and 4 objects, in which the learner has 3 extant word/object associations (See Figure 2a). Here, we are agnostic as to whether those associations were acquired in this particular episode, or whether they were brought to the episode. Words and objects are random code vectors, sampled from a multivariate Gaussian $\mathcal{N}(0,1)$. These codes are concatenated to form possible word/object combinations (See Figure 2c). We compute a linear kernel over the representations of these combinations, reflecting the covariance structure amongst the codes for different word/object pairs. The different point process models then select from the 13 remaining word/object combinations, conditioned on the 3 previous word/object combinations.

We ran 1000 simulations involving different random word and object vectors, and found that the DPP exhibits the mutual exclusivity bias 99.3 % of the time. See Figure 2. As would be expected, the PPP model randomly selects from the 13 remaining possible conjunctions. The MPP prioritizes re-use of codes across instances (re-using a word to refer to multiple objects), given that its desideratum is to select combinations *similar* to those already encountered. It has an inferential bias of "many-to-one".

Thus, when codes for combinatorial states are compositions of words and objects (keys and values), a simple algorithm that maximizes the volume spanned by the code vectors naturally produces a mutual exclusivity bias in the word learning process. Re-using either words or objects across different mappings in the same context works against maximizing the volume spanned by the vectors, as the same vector will contribute to multiple combinations.

Disjunctive Syllogism. Our second example involves the ability to make inferences like those in a classical Disjunctive Syllogism (DS) (Mody & Carey, 2016; Cesana-Arlotti et al., 2018; Pepperberg et al. 2019). Formally, a disjunctive syllogism starts with the representation of a disjunction (*premise 1: A or B*), where 'or' is XOR (one or the other, but not both). Next, one acquires some piece of information (*premise 2: not A*). Finally, a rule is applied to derive the conclusion, conditioned on the premises. (*conclusion: Therefore, B*). Notably, young

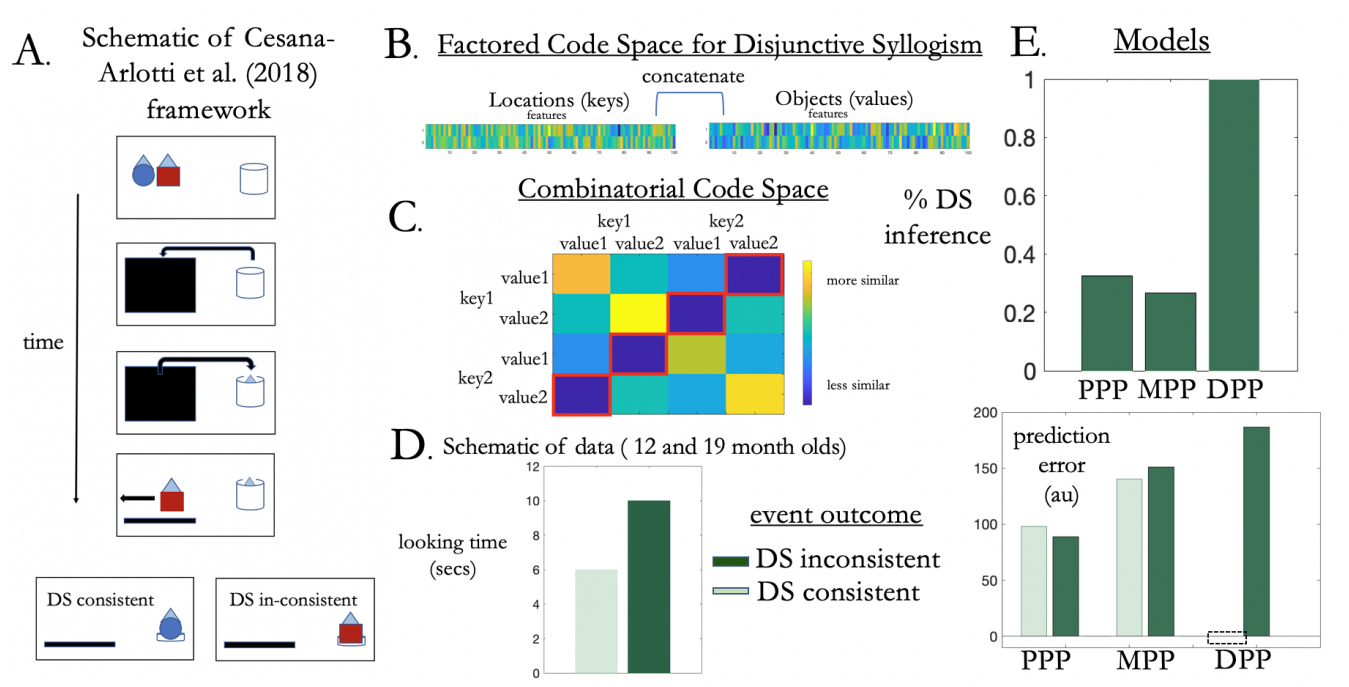

Figure 3: **(A)**. Rendering of Cesana-Arlotti (2018)'s experimental paradigm, based on their Figure 1. **(B)**. To model this, we assume a factored space of object and location codes under consideration and **(C)** populate a square matrix with the pairwise similarities between possible object/location combinations (K). We highlight in *red* the items selected by $\textit{arg max}_{S \in Y} \; det(K_S)$, conditioned separately on each combination that could be observed (the diagonal). The DPP reliably favors the combination that is most dissimilar (*dark blue*) to the observed object/location combination (*bright yellow*). **(D)**. Cesana-Arlotti et al. (2018) found that infants exhibit increased looking time to DS-inconsistent cases (results schematically depicted here). **(E)** The DPP model naturally selects a combination of the un-used location and un-used object. If these inferences are used to generate predictions and compared against the DS-consistent and DS-inconsistent cases, the DPP exhibits greater prediction errors when the revealed object/location combination is DS-inconsistent, like pre-verbal infants.

children (Mody & Carey, 2016) including infants as young as 12 months (Cesana-Arlotti et al. 2018) and non-human animals (Pepperberg et al. 2019) all show aspects of this inferential ability. Here, we show that a DPP defined over the space of combinatorial representations predicts the key empirical pattern.

For expository purposes, we focus on Cesana-Arlotti et al. (2018)'s paradigm with pre-verbal infants. See Figure 3a for a schematic of a trial. A trial begins with two objects on a screen. Both are temporarily hidden behind an occluder, obstructing the objects from the infant's view. One object is then seen to be scooped out from behind the occluder, though the infant is unable to determine which of the two objects it was. The occluder is then removed revealing (e.g.,) object A. The inference by disjunctive syllogism, of course, is that object B must therefore be the object in the bucket. Infants' expectations are assessed by measuring their looking time. If it is then revealed that the bucket contains object A, rather than object B (the "DS-inconsistent" condition), infants as young as 12 months old are surprised, evidenced by increased looking time relative to the alternative outcome in which object B is in the bucket ("DS-consistent" condition).

To model this, we assume 1x100 random code vectors drawn from a multivariate Gaussian $\mathcal{N}(0,1)$ for each of two objects (values) and two locations (keys). We concatenate these to form a 4x200 matrix, in which the rows are compositions of possible object/location combinations, and the columns are the random features. As above, we compute the covariance between each of the *mXn* combinations, here obtaining a 4x4 kernel K encoding the similarities between the codes for possible combinatorial states. For our analyses, we simulated 1000 different possible instances of random vectors, while also randomly selecting different superficial trial structures (e.g., that the DS-consistent combination was object A/location 1, object A/location2, object B/location1, objectB/location2). As expected, given one conjunction (e.g., object A in location 1), a DPP reliably selects the un-used object *and* the un-used location (here, object B in location 2), as it maximizes the volume spanned by vectors encoding the combinations. See Figure 3. To more directly relate the models' inferences to the infant looking time data, we next computed the MSE between the object/location combination selected by the model and the code for the stimulus in the DS-consistent (low-surprise)

and DS-inconsistent (high-surprise) conditions. As expected, the prediction error is high for the DPP model in the DS-inconsistent, and at zero for the DS-consistent condition. The MPP (similarity-based), and PPP (random) models do not predict this direction of the "surprisal" effect.

DPP Hashing for Collision-Free Encoding.

The empirical findings that we model here demonstrate a number of notable features of ME biases, chief amongst them: (a) they appear to be present remarkably early in development (Cesana-Arlotti et al. 2018; Lewis et al., 2020), (b) they have been observed, in different forms, across representational domains (word learning and logical reasoning about physical states), and (c) related work suggests that at least one type (DS-like inferences) may be present in some non-human species (Pepperberg et al. 2019). This raises a family of interesting theoretical questions regarding their acquisition and nature. For example, are there separate domain-specific biases, the existence of each owing to its utility in a particular domain, or do these emerge from a shared system? What are the relevant representational and inferential systems, and how are they implemented? And, of course, familiar questions regarding whether such inferential biases are acquired over phylogenetic or ontogenetic time. [2]

Although we certainly do not intend to definitively answer these questions here, we add one theoretical suggestion to the literature: we propose that ME may be a consequence of a more general strategy for efficiently encoding and retrieving unique tokens of key/value associations in memory. On such an account, it would not be surprising that ME biases are present across disparate representational domains, as long as the domain requires binding of component parts and storage for later retrieval. Moreover, it seems possible that general mechanisms for encoding and retrieving of tokens (individuals) might be be present early in development and would be employed by different species.

Why would one think that encoding and retrieving unique tokens of key/value associations in memory has anything to do with mutual exclusivity? Recall, first, that we assume that the bindings under consideration (word/object or object/location combinations) are fundamentally compositional: re-using codes to promote generalization to novel combinations. However, a compositional encoding scheme also creates the possibility of collisions between distinct instances of similar states (imagine the classic example of where you parked your car yesterday vs. two days ago). One might therefore wish to index the representations of individual instances in such a way as to avoid mapping similar states to the same address. One way to achieve this is to spread the keys broadly across the representational space (maximizing volume, as in a DPP). That is, we suggest that ME biases may exist across domains because those domains all draw on a shared memory system for associative binding, and an important feature of this system is its ability to avoid collisions by maximizing the representational "volume" of the memory keys. On this view, ME biases would arise in any representational domain that requires inference regarding novel combinations of familiar components (word/object, object/location).

To better illustrate the potential benefit of dispersing keys for efficient memory retrieval, it is instructive to consider data structures for "hashing" in computer science. A hash-function is a way of mapping from a datum to a unique index, such as a position in an array. Effective hashing seeks to avoid the time demands that are produced by sequential search techniques, which have a time-complexity of (O(n)) or binary search (O($log\ n$)). Instead, a good hash-function enables (O(1)) access times, in which readout time is invariant to the number of items in the memory. This is a classic example of a space/time tradeoff (Sedgewick & Wayne, 2011): If one is willing to expend the resources necessary to construct a vast associative array, data would almost never be mapped to the same position, and collisions would be minimized. However, this is costly in terms of space. By contrast, restricting the size of the array reduces the amount of space consumed, but risks dramatically increasing retrieval time, as one would have to search all the items in the particular location currently indexed (i.e, linear probing and chaining methods). Hashing seeks to effectively navigate this trade-off, constraining both the size of the array that is needed (space), while also minimizing the amount of computation spent resolving collisions (time). DPPs in representational space effectively avoid this tradeoff.

To see this, consider a toy case in which locations in memory are indexed by a finite set of keys, and we are able to select a key for each datum. We compare hypothetical key-selection algorithms that hash based on PPPs, MPPs, and DPPs where the latter two cases are defined over the similarity kernel for the space of key/value combinations in the dataset, as above. DPP-based key selection begins by randomly sampling a key/value pair. Then, each subsequent value in the dataset is tagged with the particular key that maximizes the total volume of the dataset of key/value pairs that have been hashed to that point (when concatenated with the value). DPP-based key selection (unlike PPP and MPP) thus implicitly discourages the re-use of keys, as this would reduce the volume of the parallelpided spanned by the code vectors for the association. Figure 4 shows the results of 1000 simulations comparing the performance of these different models. The probability of a collision in such an idealized memory system is near 0 (Figure 4a).[3] Notably, this "collision-free" property is accomplished

[2] The early-onset of the DS results of Cesana-Arlotti et al (2018) at least raise the possibility that pieces of the relevant machinery could be innate. However, recent computational modeling work from Lake (2019) provides an existence proof that an ME bias can *itself* be induced from domain-experience. Lake (2019) shows that a neural network equipped with an external memory and trained in a meta learning sequence-to-sequence paradigm can learn to apply an ME bias to novel instances of word-object pairings.

[3] We find that these infrequent collisions in the DPP can be completely eliminated by use of Pearson correlation to compute the

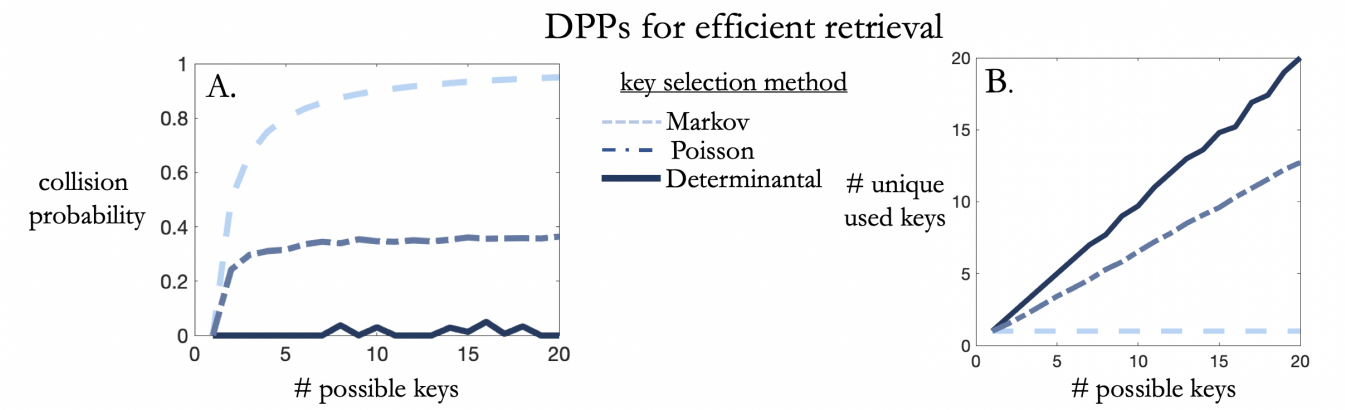

Figure 4: DPP-driven codes enable efficient retrieval of unique items. We allow the keys of key/value pairs to be selected as a MPP, PPP, or DPP. DPPs **(A)** minimize collisions between items. They do so in virtue of **(B)** selecting un-used keys to maximize the volume spanned by the key/value code vectors, across the dataset. They thus efficiently manage the resource tradeoff between search time and storage space.

without dramatically expanding the size of the array, as the array size is no larger than the number of individual states we wish to encode (Figure 4b). A DPP-based hash algorithm, unlike random selection in a PPP, thus exploits the repulsive property to distribute codes evenly across the representational space.

Discussion

We have shown that Determinantal Point Processes (DPPs)–probabilistic models of the negative interactions between states– predict a class of commonly observed biases in structured inference: specifically, inferential biases toward (a) mutual exclusivity in word learning (Markman & Wachtel, 1988; Halberda, 2003) and (b) completion of disjunctive syllogisms (Mody & Carey, 2016; Cesana-Arlotti et al., 2018). These inferences arise naturally from a DPP because a DPP selects subsets so as to maximize the volume spanned by the vectors (here, a subset of the possible combinations). When the similarity is defined over re-usable keys and values, the DPP prefers combinations of previously unused components.

This framework does not require that the cognitive system have explicit representations of rules, receive direct supervision, or have mechanisms for transferring knowledge between domains. This puts ME biases well-within the cognitive reach of pre-verbal infants, language-learning children, and non-human species that may lack the relevant experience, cortical machinery, or both necessary to represent and operate over abstract logical rules (See Mody & Carey, (2016) for related discussion regarding disjunctive syllogism in young children).

Instead, we suggest that the central driver of this bias may be the promotion of an efficient memory system. Maximizing the volume spanned by the vectors in code space promotes memory retrieval by minimizing interference ("collisions"). This is closely related to classic ideas regarding *pattern separation* in an episodic memory system (Marr, 1971; Treves & Rolls, 1994; O'Reilly & McClelland, 1994): a reduction in the similarity between two states in a function's output relative to their similarity in the input. Pattern separation is canonically implemented by projecting codes into a high-dimensional space where the probability of collisions is low. The DPP has a similar motivation here. However, a DPP precludes the need to project into a higher-dimensionality in order to avoid collisions, as it is able to uniquely map items to locations with an array size equal to the number of keys (See Figure 4). Thus, a DPP may better navigate the time/space tradeoff (Sedgewick & Wayne, 2011) than the strategy of interference-reduction through dimensionality-expansion standard in pattern separation. However, this savings in storage may come at a *computational* cost, as computing determinants has a time complexity of either $O(n^3)$ or O(!) (depending on the algorithm)[4]. These quantities therefore likely need to be approximated in order to be implemented in a neural substrate. Approximating them in a biologically plausible algorithm remains a topic of ongoing work. Although pattern separation is conventionally studied in the episodic memory literature, the theoretical points that we make throughout regarding DPPs in representation-space apply to working memory as well. In some ways, the considerations regarding structured inference are more closely to tied to what is conventionally thought

similarity matrix.

[4]We note, however, that although such exponential (or factorial) scaling is detrimental in applied use-cases of hashing, it has an intriguing connection to some empirically observed set-size effects in uniform domains, characterized by rapid, non-linear decrease in performance as n grows) (Miller, 1956; Luck & Vogel, 1997). One possibility is that capacity limits that appear to stem from a fixed number of "slots" may instead owe to the computational complexity of computing (or approximating) the determinants necessary to encode unique (non-colliding) conjunctions. At the moment, this remains speculative, however.

of as "working memory", as we assume that attentional mechanisms have already windowed into a smaller region of the possibility space so that we can easily compute the MAP over a sub-set of the broader set of possible items. Better understanding how DPPs may relate to the particular factorizations of memory systems standard in cognitive science (e.g., episodic / working / semantic), as well as specific aspects of the entorhinal/hippocampal system[5] also remain important topics of ongoing work. For present purposes, however, the central distinction in memory systems is simply that between a stable set of re-usable representations and combinations of those representations in particular instances (key/value pairs). While re-using codes promotes generalization, it increases the risk of collisions in the memory system. Here, we have suggested that an algorithm that seeks to maximize the total volume of the constructed combinations in a representational-space (exhibiting repulsion) not only promotes efficient memory encoding and retrieval, but may also underlie inferential biases toward mutual exclusivity.

Acknowledgments

This project / publication was made possible through the support of grants from the John Templeton Foundation and NIH grant T32MH065214. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509.

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, *359*(6381), 1263–1266.

Chanales, A. J., Oza, A., Favila, S. E., & Kuhl, B. A. (2017). Overlap among spatial memories triggers repulsion of hippocampal representations. *Current Biology*, *27*(15), 2307–2317.

Gillenwater, J., Kulesza, A., & Taskar, B. (2012). Near-optimal map inference for determinantal point processes. In *Advances in neural information processing systems* (pp. 2735–2743).

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Kulesza, A., Taskar, B., et al. (2012). Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, *5*(2–3), 123–286.

Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. In *Advances in neural information processing systems* (pp. 9788–9798).

Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*.

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121–157.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *262*(841), 23–81. Retrieved from http://www.jstor.org/stable/2417171

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*(3), 419.

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the society for research in child development*, i–129.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40–48.

O'reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682.

Pepperberg, I. M., Gray, S. L., Mody, S., Cornero, F. M., & Carey, S. (2019). Logical reasoning by a grey parrot? a case study of the disjunctive syllogism. *Behaviour*, *156*(5-8), 409–445.

Sedgewick, R., & Wayne, K. (2011). *Algorithms*. Addison-wesley professional.

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*(3), 374–391.

Whittington, J., Muller, T., Mark, S., Barry, C., & Behrens, T. (2018). Generalisation of structural knowledge in the hippocampal-entorhinal system. In *Advances in neural information processing systems* (pp. 8484–8495).

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: building memories to navigate future decisions. *Frontiers in human neuroscience*, *6*, 70.

[5]See Chanales et al. (2017) for intriguing fMRI evidence of "repulsion" in hippocampal codes, in which similar states have dissimilar representations.