Journal of Mathematical Psychology 109 (2022) 102675

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Theoretical Note

Clustering and the efficient use of cognitive resources

Ishita Dasgupta [a,*], Thomas L. Griffiths [b]

[a] Department of Computer Science, Princeton University, United States of America
[b] Departments of Psychology and Computer Science, Princeton University, United States of America

ARTICLE INFO

Article history:
Received 17 November 2021
Received in revised form 29 April 2022
Accepted 10 May 2022
Available online xxxx

Keywords:
Bayesian inference
Resource rationality
Information theory
Probabilistic numerics

ABSTRACT

A central component of human intelligence is the ability to make abstractions, to gloss over some details in favor of drawing out higher-order structure. Clustering stimuli together is a classic example of this. However, the crucial question remains of how one *should* make these abstractions—what details to retain and what to throw away? How many clusters to form? We provide an analysis of how a rational agent with limited cognitive resources should approach this problem, considering not only how well a clustering fits the data but also by how 'complex' it is, i.e. how cognitively expensive it is to represent. We show that the solution to this problem provides a way to reinterpret a wide range of psychological models that are based on principles from non-parametric Bayesian statistics. In particular, we show that the Chinese Restaurant Process prior, ubiquitous in rational models of human and animal clustering behavior, can be interpreted as minimizing an intuitive formulation of representational complexity.

1. Introduction

The only enduring aspect of our environment is that nothing stays the same. We never have *exactly* the same experience twice. As a consequence, the human mind has to form abstractions, clustering these experiences together in a way that supports generalization. Psychological models have applied this lens to phenomena as diverse as categorization (Anderson, 1991; Sanborn et al., 2010), feature learning (Griffiths & Austerweil, 2008), theory formation (Kemp et al., 2010), classical conditioning (Gershman et al., 2010), and word segmentation (Goldwater et al., 2009). A key problem that arises in each of these models is knowing when to generate a new cluster—when an object, stimulus, or word is genuinely of a kind that has never been seen before.

Deciding when to form a new cluster involves making a tradeoff between the complexity of the underlying representation and how well it describes the environment. Grouping all experiences into a single cluster where they are represented by some abstract summary statistics is maximally simple, but at the cost of losing a significant amount of detail. Having a separate cluster for each experience accurately captures the nuances of those experiences, but is maximally complex. So how *should* we form clusters?

In this paper, we address this question in the spirit of rational analysis (Anderson, 1991), asking how it might be solved by an ideal agent. More precisely, we engage in resource rational analysis (Gershman et al., 2015; Griffiths et al., 2015; Lieder & Griffiths, 2019), since our analysis focuses on the question of how that agent might make the best use of limited cognitive resources. Formalizing the complexity of a clustering in information-theoretic terms, we derive an optimal distribution over clusterings.

This analysis yields a surprising result: our optimal solution has the same properties as the distribution over clusterings assumed in *all* of the psychological models mentioned above. These models use a distribution over clusterings originally introduced in psychology by Anderson (1991) in his rational model of categorization. This distribution was independently discovered in nonparametric Bayesian statistics (Aldous, 1985; Hjort et al., 2010), where it is known as the Chinese restaurant process (CRP).[1]

The CRP has a number of attractive mathematical properties that can be used to justify its use in psychological models, related to its convenience or the assumptions it makes about the environment. Our analysis provides a new reason why the CRP might make sense as a component of psychological models: we show that CRP-like distributions can arise from an effort to minimize representational costs, i.e. that this distribution is normative under an assumption of resource-rationality. In particular, we show that best using a fixed number of bits to store an object-category mapping (viz. controlling the complexity of that mapping) can

* Correspondence to: DeepMind New York City, USA.
E-mail address: idg@google.com (I. Dasgupta).

[1] The name of this process comes from its creators imagining a large restaurant that seats multiple parties at communal tables, with people joining tables based on their current popularity – a phenomenon that could apparently be observed in San Francisco's Chinatown. The tables are clusters and the people the experiences being clustered.

https://doi.org/10.1016/j.jmp.2022.102675
0022-2496/© 2022 Elsevier Inc. All rights reserved.

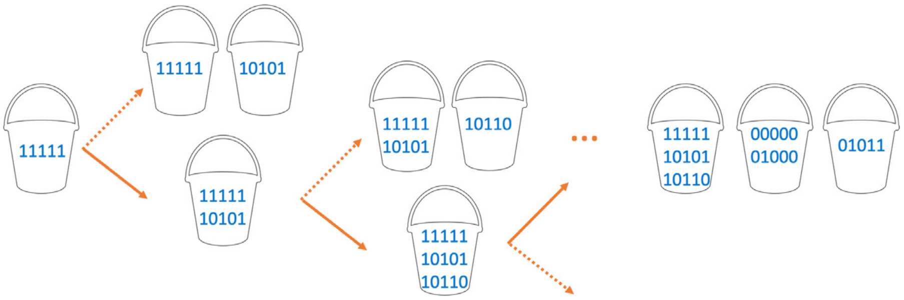


Fig. 1. Schematic illustration of the categorization task simulated in Anderson (1991). In this figure, we represent each object with its binary features and each cluster as a bucket. Each time an object is added, we can either assign it to an existing cluster, or create a new one. Solid arrows indicate the higher probability assignment, Anderson (1991) assumes that this is the assignment chosen at each step. Iterating this over the 6 stimuli in their (fixed) order of presentation gives the final clustering. See main text for further details.

give CRP-like behavior. These results provide the first process-level explanation of why this kind of clustering behavior might be a reasonable assumption in psychological models.

The plan of the paper is as follows. In the next section we provide a more detailed introduction to the Chinese restaurant process. We then turn to our analysis of optimal clustering under resource constraints. We derive our key results mathematically and present simulations that verify our analysis. We conclude with a discussion of the implications of these results for developing models of human cognition.

2. Background: The Chinese restaurant process

As mentioned above, one of the challenges involved in clustering a set of experiences is deciding how many clusters there should be. Researchers in nonparametric Bayesian statistics developed an innovative strategy for solving this problem: rather than specifying a particular number of clusters, we instead assume that there could exist an infinite number of clusters of which only a finite number have been observed so far. The problem of determining the number of clusters then becomes a matter of inferring how many clusters may have been observed, which can be solved by applying Bayes' rule.

Pursuing this approach requires identifying a prior distribution over clusterings that remains well-defined regardless of how many experiences need to be clustered. A common way to achieve this goal is to assume that the prior probability an item belongs to a cluster follows a distribution known as the Chinese Restaurant process (CRP; Aldous, 1985). Under this distribution, the probability of belonging to an existing cluster is proportional to the number of objects already in that cluster, while the probability of a new cluster is proportional to the value of a parameter α.

To make this kind of Bayesian clustering more concrete, we consider how it might apply to an empirical context: the first experiment from Medin and Schaffer (1978). Here, participants saw and were asked to categorize stimuli that varied along four binary dimensions (color, form, shape and position) and given a binary category label. These data were used to demonstrate the rational model of categorization presented in Anderson (1991) (illustrated in Fig. 1), which assumes a prior on clusters that is equivalent to the CRP. Anderson assumed that the binary category label is treated as an additional feature, so there are five binary features. The stimuli are then assumed to be presented in the order $\{11111, 10101, 01011, 00000, 01000, 10110\}$. Anderson's model assigns the first stimulus to its own cluster. On seeing the second stimulus, it decides whether to assign it a new cluster or to the same cluster as the first one. It makes this decision on the basis of two factors: the new stimulus' overlap with the features of the stimuli already in a cluster (the likelihood function), and a parameter that determines how likely in general it is that two stimuli belong to the same cluster vs. different ones (the prior). The assumption that Anderson (1991) made about the latter parameter defines a prior that is equivalent to the CRP (Neal, 1998).[2]

Iterating this process for each stimulus, Anderson's model predicted that the most likely clustering of these stimuli is $(11111, 10101, 10110), (00000, 01000), (01011)$. This clustering does not split the stimuli by the category label (here, the last binary feature). However, the model can generate predictions for the category membership of novel stimuli based on this clustering by calculating the probability they get assigned to a cluster and the probability of the category label under that cluster. Anderson showed that these predictions are in close accordance with the judgments of participants in Medin and Schaffer's (1978) experiment, with a rank-order correlation of .87.

Anderson's model used a particularly simple inference algorithm for the CRP, allocating each stimulus to the cluster with highest posterior probability based on previous allocations (as also depicted in Fig. 1). Subsequent work has extended this model to accommodate different inference algorithms (Sanborn et al., 2010) and sharing of clusters across categories (Griffiths et al., 2007), applying the resulting models to a variety of results in human category learning (for a review, see Griffiths et al., 2008).

The CRP has various desirable properties that make it a sensible choice as a prior over clusters. In the infinite space of possible clusterings, it favors assigning data to a small number of clusters. The expected number of clusters grows slowly as the number of experiences being clustered increases. In particular, the CRP displays "preferential attachment" or a "rich-get-richer" dynamic, where a cluster with a large number of members is more likely to grow further. The resulting distributions of cluster sizes ('scale-free' distributions, where the size of clusters decays as a power law) have been shown to be prevalent across several other domains (Adamic & Huberman, 2002; Barabási & Albert, 1999; Mandelbrot, 1960; Rosen & Resnick, 1980).

Another practical reason for the success of the CRP is that it is agnostic to the order of data presentation (i.e. exchangeable; Aldous, 1985)—changing the order of presentation of experiences does not change the probability of their cluster memberships. This makes Bayesian inference more tractable, as it is easy to compute conditional distributions that are required for standard inference algorithms (see, e.g., Gershman & Blei, 2012).

[2] Intuitively, with a fixed probability that two objects belong to the same cluster (c, the *coupling constant* in Anderson (1991)), it is more likely that the new object should be in the same cluster as an object that already has several objects in it. This results in the CRP assignment rules outlined above, with the coupling constant scaling inversely with α: $\alpha = (1 - c)/c$.

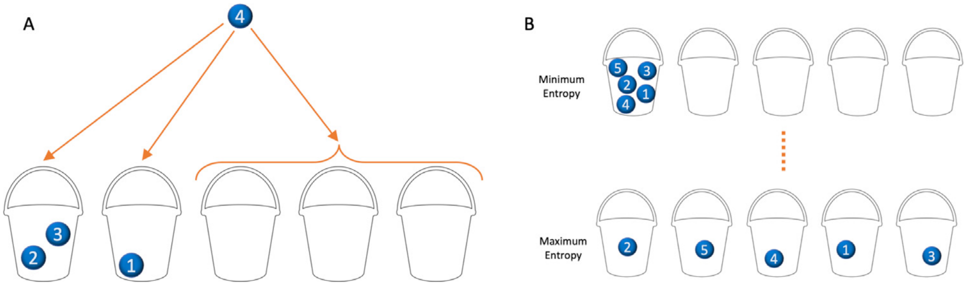


Fig. 2. Schematic illustration of the clustering problem. In this schematic, we represent each object as a blue ball and each cluster as a bucket. Each object has no features except its unique index. (A) We want to cluster object 4, conditioned on how objects 1, 2, and 3 were clustered. There are three possibilities (indicated by the orange arrows): assign to bucket 1 which contains two objects already, assign to bucket 2 with one object, or start a new cluster by assigning to one of the empty buckets. In this work, the prior over clusterings specifies these conditional probabilities. (B) By iterating these conditional probabilities, we can derive a probability distribution over the range of possible final clusterings of all objects (in this case 5 objects in total). These final clusterings are represented here, as well as how they differ in the entropy of the marginal distribution over clusters. A prior over clusterings specifies a distribution over these different solutions.

In addition to these practical properties of the CRP, there are other reasons why human minds might use this particular prior distribution for clustering. In its first use in psychology, Anderson (1991) derived the CRP from the assumption that any two objects must have the same fixed prior probability of being in the same cluster. This is related to exchangeability, and might be an assumption justified by the environment in which humans operate. In the remainder of the paper we pursue another hypothesis about the appropriateness of the CRP for cognitive modeling: that CRP-like clustering (as well as its various properties, like exchangeability) might be emergent properties of a resource-rational tendency to best utilize a limited representational budget.

3. Resource-rational clustering

As noted above, the CRP has been used to model clustering problems that arise in a variety of domains. Following Anderson's (1991) original application we will focus on the case where the agent seeks to organize a set of objects into clusters to support their categorization (see Fig. 2). We formalize this problem as follows. We have a total of N objects. Since we are concerned with examining the *prior* over clusterings (i.e., how each object should be assigned to a cluster in the absence of any specific features), we assume that these objects do not have any distinguishing features except for their index $i \in [0, N]$. The goal is to organize these objects into clusters. We do not know a priori how many clusters the objects will be sorted into, but they will certainly be no more by the number of objects N. We therefore need to learn a mapping π from object o_j for $j \in [0, N)$ to cluster c_i for $i \in [0, N)$.

For an agent with finite cognitive resources, it will be important to represent these objects in as simple a way as possible while allowing for the potential differences between them. We will derive a prior based on this idea of learning 'simpler' mappings π, and show that this simplicity prior corresponds closely to the CRP. But first, how do we measure simplicity?

3.1. Measuring simplicity or complexity

A human preference toward simplicity, or Occam's razor, has been used to explain several cognitive phenomena in perception, learning and high-level cognition (Chater & Vitányi, 2003), with the use of information theory to define this 'simplicity' (e.g., Bhui & Gershman, 2018; Gottwald & Braun, 2019; Olshausen & Field, 1996; Ortega et al., 2015; Todorov, 2009; Zenon et al., 2019). We follow in this tradition and use the length of the code required to represent a mapping π as a measure of its complexity (as also seen in Chater, 1996; Feldman, 2016). A longer (i.e. more complex) code has higher representational cost. We make this more precise below.

We first compute an intermediate quantity, the marginal distribution over categories given a mapping π:

$$P_\pi(c_i) = \frac{\sum_j^N \mathbb{I}[\pi(o_j) == c_i]}{N} \tag{1}$$

Each mapping π gives a probability distribution over categories. We then define simplicity or complexity of this probability distribution. What makes one distribution over clusters more or less complex than another?

The entropy of the distribution can act as a measure of its representational cost and thereby of its complexity. It is given by:

$$\mathcal{H}(\pi) = -\sum_i^N P_\pi(c_i) \log P_\pi(c_i) \tag{2}$$

The information-theoretic interpretation of the entropy of a distribution is that it measures the average number of bits (binary coin flips) required to convey an object sampled from that distribution, under the most efficient code possible. The number of bits required for c_i, or the length of its 'codeword', is $\log P_\pi(c_i)$ (under the most efficient code; Shannon, 1948). Weighting this codeword length by the probability of each token gives the entropy of the distribution. Intuitively, this measures how difficult it is (i.e. how many bits of information are required) to convey which object is sampled, when randomly sampling objects from the given distribution, to an observer that knows the distribution but does not know which specific object was sampled. A representationally 'expensive' or 'complex' distribution is one that requires more such bits.

In using entropy as a measure of representational complexity, we are following previous work in both psychology and neuroscience. Work on planning and sequential decision-making has used entropy as a measure of representational cost (Todorov, 2009), and other work has suggested that a tendency to minimize this information-theoretic cost is what characterizes bounded-rational behavior in agents with limited resources (Olshausen & Field, 1996; Ortega et al., 2015). This tendency has been empirically validated, and used to model neural representations (Laughlin, 1981) as well as human behavior in high-level cognitive tasks (Bhui & Gershman, 2018).

How does this measure map onto our intuitions in this domain? The lowest entropy distribution is the distribution that allocates all of its probability to a single outcome. Here, the entropy is zero, since samples from the distribution are always the same—there is no information to be transmitted about a specific sample. On the flip side, the highest entropy distribution is one that is uniform over all outcomes. Here, since all outcomes are equally likely, even the most efficient code has to convey which of several possibilities was actually chosen at a given sample. Other distributions fall in between, as measured by Eq. (2). In the context of our clustering problem, this maps onto the intuition that it is easy to remember to always put every object in the same cluster (a low entropy distribution, lower representational cost), but harder to remember different clusters for each object—with the extreme being to have a separate cluster for each object (a high entropy distribution, higher representational cost).

We want to use this measure of complexity to inform a probability distribution $P(\pi)$ over mappings π. We do have some representational capacity and would like to best utilize it—we do not want to simply minimize the entropy and default to the zero entropy π. Instead we assume some fixed average representational capacity—this means that the mean entropy averaged over all clusterings (weighted by $P(\pi)$) is fixed. Since the number of possible clusters is infinite, the entropy can grow arbitrarily large. The probability of these high entropy distributions must be correspondingly low to accommodate finite representational capacity in expectation. We therefore will prefer low complexity (low entropy) mappings over higher complexity ones. Exactly how should this preference decay as a function of complexity/entropy? The negative exponential is the maximum entropy distribution for a fixed mean (note that 'entropy' in 'maximum entropy' here refers to the entropy of the prior probability distribution $P(\pi)$—not $\mathcal{H}(\pi)$ which is the entropy of the mapping π). This gives:

$$\begin{aligned} P(\pi) &= \frac{\exp(-k\mathcal{H}(\pi))}{\sum_{\pi'} \exp(-k\mathcal{H}(\pi'))} \\ &\propto \exp(-k\mathcal{H}(\pi)) \end{aligned} \tag{3}$$

where k is a positive constant and the normalizing factor sums over all possible mappings π'. It is (by the principle of maximum entropy; Shore & Johnson, 1980) the most general, least informative distribution given a fixed mean, i.e. a fixed average representational capacity. This is the same logic used to derive other assumption-free distributions with fixed mean e.g. the negative exponential free energy functional (Ortega et al., 2015).

We have therefore specified a prior probability distribution $P(\pi)$ over different clusterings π for when we have a fixed mean representational capacity, where representational cost of a mapping is given by $\mathcal{H}(\pi)$. In the following sections, we show that the CRP corresponds to exactly such a probability distribution.

3.2. The relationship between the CRP and entropy

We first discuss the key properties of the CRP. The key property of the CRP is the way a new object is added to an existing clustering of states:

1. Assign it to an existing cluster with probability proportional to the number of items already in the cluster.
2. Assign it to a new cluster with fixed probability α.

This "rich-get-richer" or "preferential attachment" also arises when trying to reduce entropy. Adding an object to a cluster that already has high probability reduces the entropy of the distribution by making it peakier. On the other hand, adding it to a less populated one moves the distribution closer to uniform, increasing its entropy. Therefore, adding a new object to a cluster that already has many objects in it results in less cost for representing that distribution than adding it to one that has fewer objects.

We can formalize this intuition. By Eq. (3), the entropy of a mapping specifies its prior probability. We can compute the entropies of all the mappings that arise from different possible assignment of a new object to an existing mapping. Inserting these entropies into Eq. (3) specifies a probability distribution over the possible assignments of the new object. This way of assigning new objects to clusters is not arbitrary. Rather, it is normative, under the resource-rational assumption that we want to best utilize limited representational resources. We thereby prefer mappings with low complexity (and hence low representation cost), see Section 3.1 for details.

In the following section, we provide the mathematical details of the above procedure.

4. A mathematical derivation

In this section, we show that assigning new objects based on probability under Eq. (3) recovers the CRP's new object assignment rules when the number of objects being classified is reasonably large.

4.1. Conditional distributions of the CRP and entropy-based priors

Given a cluster assignment π, an object is added to give π'. This can be split into two cases, where the object is either added to cluster j to give π_j or it is added to a new cluster (i.e. one with $n_j = 0$) to give π^0. Formalizing the CRP in these terms:

$$\begin{aligned} &P_{crp}(\pi_j|\pi) \propto n_j \\ &P_{crp}(\pi^0|\pi) \propto \alpha \end{aligned} \tag{4}$$

The entropy-based prior over possible cluster assignments (as given in Eq. (3)) is proportional to the negative exponent of the entropy of the mapping π. The equivalent conditional distributions for this prior are given by (details in Appendix A):

$$\begin{aligned} P_{entropy}(\pi_j|\pi) &\propto \exp\left(-k(\mathcal{H}_j - \mathcal{H})\right) \\ P_{entropy}(\pi^0|\pi) &\propto \exp\left(-k(\mathcal{H}^0 - \mathcal{H})\right) \end{aligned} \tag{5}$$

where the entropy of a mapping (specifying a cluster assignment of N objects, with each cluster containing n_i objects) is given by:

$$\mathcal{H} = -\sum_{i}^{N} \frac{n_i}{N} \log\frac{n_i}{N} \tag{6}$$

In the next section, we compute the differences $\mathcal{H}_j - \mathcal{H}$ and $\mathcal{H}^0 - \mathcal{H}$ in terms of n_j to more directly compare the conditionals derived from the CRP and the entropy-based prior.

4.2. The effect of adding an object on entropy

When adding a new object, we can add it to an existing cluster j to give $\mathcal{H}_j$.

$$\mathcal{H}_j = -\sum_{i \setminus j}^{K} \frac{n_i}{N+1} \log\frac{n_i}{N+1} - \frac{n_j+1}{N+1} \log\frac{n_j+1}{N+1}$$

We take the difference with $\mathcal{H}$ in Eq. (6) and separate out the terms independent of n_j (denoted E) from those dependent on n_j. This simplifies to the following (see Appendix B for detailed steps):

$$\mathcal{H}_j - \mathcal{H} = E - \frac{n_j \log\left(1 + \frac{1}{n_j}\right)}{N+1} - \frac{\log(n_j+1)}{N+1}$$

We then take the large N limit and consider only the leading order terms. To decide which of these terms are leading, we need to make an assumption about the relation between the number of objects in a cluster (n_j) and the total number of objects (N). We make the weak assumption that the average number of objects in a cluster grows sub-linearly with the total number of objects—this holds as long as (a) not all objects are assigned to the same clusters or (b) not all objects are assigned to a new cluster. In other words, both n_j and N grow when N is large, but n_j grows slower. We can therefore Taylor expand the first term and keep only leading order terms. This simplifies to the following, detailed steps in Appendix B:

$$\mathcal{H}_j - \mathcal{H} \sim \; E - \frac{\log(n_j)}{N} - \frac{1}{N} \tag{7}$$

We have so far derived the change in entropy when we add an object to an existing cluster with n_j objects. The change to entropy when we instead add an object to an empty cluster is given by $n_j = 0$ (before applying any approximations in the large N limit):

$$\mathcal{H}^0 - \mathcal{H} = E$$

4.3. From entropy to probability distribution

Substituting $\mathcal{H}_j - \mathcal{H}$ and $\mathcal{H}^0 - \mathcal{H}$ into the expressions for the conditional distributions of the entropy-based prior (Eq. (5)) gives:

$$\begin{aligned} P(\pi_j|\pi) &\propto \; \exp\left(-k(\mathcal{H}_j - \mathcal{H})\right) \\ &= \exp\left[-k\left(E - \frac{\log(n_j)}{N} - \frac{1}{N}\right)\right] \end{aligned}$$

We fix $k = N$; this constant k (from Eq. (3)) is therefore not a parameter.

$$\begin{aligned} P(\pi_j|\pi) &\propto \exp(-NE) \times n_j \times e \\ &\propto n_j \end{aligned}$$

Similarly, for when we are adding to a new cluster:

$$\begin{aligned} P(\pi^0|\pi) &\propto \; \exp\left(-N(\mathcal{H}^0 - \mathcal{H})\right) \\ &= \exp\left(-NE\right) \end{aligned}$$

We can then normalize the probability of the new clusterings as follows:

$$P(\pi_j|\pi) = \frac{P(\pi_j|\pi)}{P(\pi^0|\pi) + \sum_i P(\pi_i|\pi)} = \frac{n_j}{N + 1/e}$$

$$P(\pi^0|\pi) = \frac{P(\pi^0|\pi)}{P(\pi^0|\pi) + \sum_i P(\pi_i|\pi)} = \frac{1/e}{N + 1/e}$$

This is equivalent to a CRP (as specified in Eq. (4)) with $\alpha = 1/e \approx 0.36$. Note that we can get a corresponding CRP with a different α by taking the logarithm in Eq. (2) and the exponent in Eq. (3) with respect to a different base. The base of the logarithm is restricted to be greater than 1, to ensure that the logarithm is an increasing function, but is arbitrary beyond this constraint. We can thus derive the full family of CRP distributions for $0 < \alpha < 1$.

4.4. Summary

Our goal was to examine the consequences of limited cognitive resources on the clustering process. We find that a prior over clusters proportional to the negative exponent of the entropy of the cluster mapping gives CRP-like clustering. This prior strongly prefers lower entropy mappings to higher entropy ones. This indicates that CRP-like clustering might come from a tendency to reduce representational burden. In other words, CRP-like clustering can come from a bias toward learning 'simple' object-category mappings, where simplicity is defined as the entropy of the marginal over categories. This provides a process-level theory for why CRP-like priors might be appropriate for modeling human cognition.

5. Simulations

Our mathematical results establish a direct correspondence between limited representational capacity and the CRP, in the limits of a large number of objects. To determine whether the clustering produced by this resource-rational clustering scheme produces results similar to those expected from a CRP with realistic samples, we conducted a series of simulations where we generated cluster assignments for both distributions and then analyze the correspondence.

The correspondence between the CRP and the entropy-based distribution is closer as the number of objects (N) increases. At very low N therefore, these distributions deviate slightly (see Appendix C for details). Since subsequent clustering behavior is conditioned strongly on the object assignments thus far, these differences can amplify. That is, even though the conditional distributions get closer with higher N, the marginal distributions deviate further. The resultant distribution is still qualitatively very similar to the CRP (as discussed below), and it is an interesting direction of future work to examine whether this distribution might *better* describe human clustering behavior than traditional CRPs. Here, however, to validate the similarities with the CRP, we control for this deviation by clustering the first M objects according to an exact CRP.

We evaluated the correspondence between the CRP and our resource-rational distribution based on two criteria. First, a property characteristic of scale-free distributions like the CRP is that the sizes of the different clusters decay as a power law. Therefore, if we sort the clusters by size, the logarithm of the cluster size (i.e. the fraction of the total number of objects that are in that cluster) should be a linear function of the logarithm of the cluster index. Second, another key property of the CRP is that the number of clusters increases logarithmically with the total number of objects. We can also measure that for our entropy-based distribution and examine whether the number of clusters is a linear function of the logarithm of the number of objects.

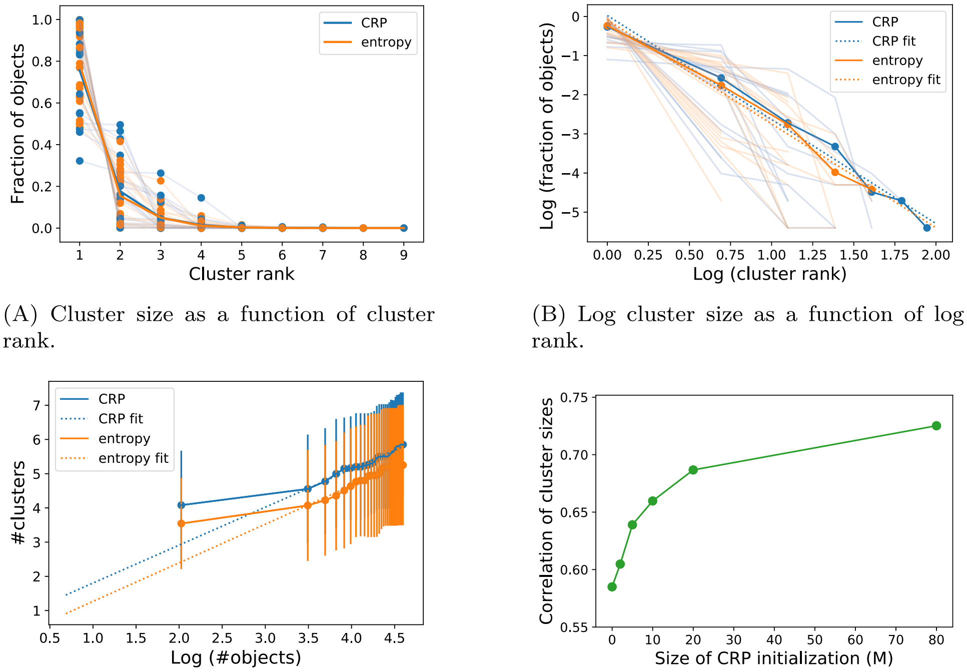
5.1. Method

We generated samples from our distribution as follows. We first cluster M objects according to a CRP with $\alpha = 0.368$, varying M between 0 and 80. We then cluster an additional 10^6 objects from this starting point, either with the CRP's conditional cluster assignment rules, or based on the entropy as specified by Eq. (3). We cluster such a large number of objects to get a reasonable number of total clusters so that we can better analyse the distributions of objects across them—even with these many objects, the average number of total clusters is around 6. This procedure is repeated to get 20 unique set of clusters for each cluster assignment rule and each M.

5.2. Results

We find that the entropy-based distribution produces clusterings that have statistical properties that closely resemble those from the CRP. To provide an initial illustration of the correspondence, we focus on the case where $M = 20$. Fig. 3A shows how the size of the clusters decays as a power law over the cluster index ranked by size. Fig. 3B plots these in log-space and highlights the linear relationship characteristic of a power law like the CRP.

It is especially instructive to compare this linear pattern with what might be expected from other sensible priors over clusters. For example, ones where the cluster size decays exponentially with rank (intuitively, each cluster is some fixed factor λ smaller than the next largest one). This prior notably does not show this



(A) Cluster size as a function of cluster rank.

(B) Log cluster size as a function of log rank.

(C) Number of clusters vs. log number of objects.

(D) Correlation in cluster sizes with increasing M.

Fig. 3. Simulation results: (A) Cluster size scales as a power law with cluster rank. (B) This relationship is highlighted in log space. The average behavior of the two clustering assignments resembles the linear fit (dotted lines). (C) The number of clusters grows logarithmically with the number of objects. (D) The correlation between the cluster sizes from the two clusterings increases with M, but levels off.

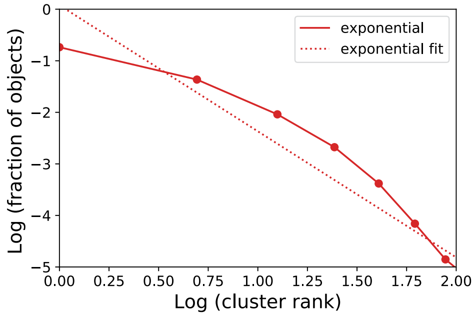


Fig. 4. Alternative priors: Log cluster size as a function of log rank for an exponential prior over ranked clusters. The dotted line shows a linear fit to the data: this fit does not explain this data well. In comparison, Fig. 3B shows that the entropy-based prior follows the same linear trend as the CRP.

linear structure (as plotted in Fig. 4). It is therefore particularly remarkable that our entropy-derived prior also shows this linear trend.

Fig. 3C shows the change in the average number of clusters (over the 20 runs) as a function of the logarithm of the number of objects. We see that this grows close to linearly, with the linear fit (dotted line) closely matching the data. The deviation is most apparent at smaller numbers of objects, as expected since the linear relationship is expected in the limit. The number of clusters from the CRP is slightly but not statistically significantly higher than that from the entropy-based clustering.

To provide a more detailed picture of this correspondence, we examined the correlation between the cluster sizes (i.e. the data plotted in Fig. 3A, matched by cluster index) for the first 500 objects. Cluster assignment is stochastic and we cannot expect exact correspondence. To get a sense of the upper limit of the correlation we can expect from this measure, we first correlate the cluster sizes derived from independent runs of the same clustering algorithm, repeated 4 times. This gives a correlation of 0.72 (95% CI [0.69, 0.74]) for the CRP, and 0.74 (95% CI [0.71, 0.76]) for entropy-based clustering with $M = 80$. This shows that even with the cluster sizes produced by the same algorithm, we cannot expect a correlation of 1.

We then look at the correlation between the CRP and the entropy-based clustering after initializing with variable M. We expect this correlation to improve as M increases. We see that this is indeed the case (Fig. 3D), with the correlation at $M = 80$ being comparable to the correlation between two runs of the same clustering algorithm (CRP or entropy-based). We also see that the difference in correlation from small to large M is not very dramatic (varies from ∼0.58 to ∼0.73) and appears to level off. This indicates that the correspondence between the CRP and the entropy-based prior is fairly robust to the value of the initialization M.

6. Discussion

Needing to cluster experiences together is a ubiquitous aspect of human cognition. In this paper, we have approached this problem from the perspective of rational analysis, asking how an ideal agent should seek to use their limited cognitive resources. Our results show that the solution to this problem, when those resources are expressed in information-theoretic terms, has a direct correspondence with an approach to clustering that has been widely used in probabilistic models of cognition (the Chinese restaurant process, or CRP). These results provide a new cognitively-motivated justification for that assumption.

Our findings suggest interesting directions for future empirical work. If CRP-like clustering comes from representational costs, manipulating these costs should result in different clustering behaviors. Our model predicts that having more limited cognitive resources should affect clustering behavior, driving toward a lower entropy representation and a stronger preference for few, large, clusters. Gershman and Cikara (2021) model the effects of cognitive load on structure learning as a reduction in the concentration parameter giving fewer clusters. Our approach provides theoretic justification for why fewer cognitive resources (e.g. under cognitive load) should give rise to fewer clusters. This would not be predicted by a traditional CRP model, since it is a consequence of the cognitive resources available and not a change in the beliefs of the agent about the relative prior probability of different clusterings.

In our paper, we do not assume that the data are generated from a ground-truth set of clusters, rather that clustering arises solely at the representational level from limitations in capacity. Correspondingly, we make no assumptions about the likelihood function that informs within-cluster structure – we focus entirely on the a priori number of clusters, assuming the data have no distinguishing features to cluster on the basis of. However, these are crucial aspects of real-world clustering behavior and future work should look toward how they interact with a priori clustering driven by representational limits as posited here. A common criticism of Bayesian models of cognition is their lack of grounding in process-level considerations, and the risk that the priors specified in these models can be arbitrarily chosen by practitioners to fit data (Bowers & Davis, 2012; Jones & Love, 2011). Our work exemplifies one way to specify 'effective priors' that are informed and constrained by algorithmic considerations—and in fact directly arise from resource limitations at this algorithmic level. Further, posterior inference with arbitrary priors is often computationally intractable; this approach also suggests a tractable process-level model. This raises the broader question of the epistemic role of the prior in Bayesian models of cognition—whether it represents pre-existing knowledge, or emergent properties of the algorithm. Our work highlights that this difference can be nuanced.

An open question is whether other ubiquitous priors assumed in probabilistic models of human cognition might instead arise from the algorithmic processes involved in learning and representation. Various priors over neural network models have been shown to be effectively implemented by established algorithmic approaches like weight decay (Krogh & Hertz, 1991), early stopping (Duvenaud et al., 2016), and dropout (Gal & Ghahramani, 2016). The field of probabilistic numerics (Hennig et al., 2015) has also shown that several classic approximate algorithms in quadrature, linear optimization, and solving differential equations can be reinterpreted as exact solutions under specific priors. These approaches (e.g. neural networks, linear optimization) are commonly used in probabilistic models of cognition. Exploring this duality (between algorithmic process and computational prior) within these approaches – and therefore the role these approaches play in modeling cognition – is a promising direction of future work.

Acknowledgments

This work was done while ID was at the Department of Computer Science, Princeton University. We are grateful to Michael Chang, Jon Cohen, Zack Dulberg, Noémi Éltető, Sam Gershman, Mark Ho, Eric Schulz, Simon Segert, and Shuchen Wu for helpful discussions. This work was supported by DARPA, United States of America grant FA8650-18-2-7832 and AFOSR, United States of America grant FA9550-18-1-0077.

Appendix A. Conditional distributions for the entropy-based prior

We give further detail on how to derive the conditional distributions from the negative exponential probability over mappings π. We start with:

$$P(\pi) = \frac{\exp(-k\mathcal{H}(\pi))}{\sum_{\pi'} \exp(-k\mathcal{H}(\pi'))} \propto \exp(-k\mathcal{H}(\pi))$$

Adding an object o_i to cluster j gives π_j. The probability of this new mapping is given by $P(\pi_j) \propto \exp(-k\mathcal{H}(\pi_j))$. We can then consider the probability of $P(\pi_j)$ conditioned on already having the mapping π.

$$P(\pi_j|\pi) = \frac{P(\pi_j, \pi)}{P(\pi)}$$

Note that π_j specifies a super set of π, i.e. π specifies the cluster mapping of objects $o_0 \dots o_{i-1}$, while π_j additionally specifies the mapping of o_i. So $P(\pi, \pi_j) = P(\pi_j)$.This reduces the conditional distribution to

$$P(\pi_j|\pi) = \frac{P(\pi_j)}{P(\pi)} \propto \exp(-k(\mathcal{H}(\pi_j) - \mathcal{H}(\pi)))$$

In terms of CRP notation that directly represents the probabilities of the cluster assignments $z_i = \pi(o_i)$, we have $P(z_i = j|z_1, \dots, z_{i-1}) = P(\pi_j|\pi)$.

Appendix B. Simplifying the difference in entropy

The difference in the entropy between the new and the old distributions is:

$$\begin{aligned}\mathcal{H}_j - \mathcal{H} = &\sum_{i \backslash j}^{K} n_i \left(\frac{1}{N} - \frac{1}{N+1}\right) \log n_i \\ &+ \frac{n_j}{N} \log n_j - \frac{n_j+1}{N+1} \log(n_j+1) \\ &+ \log(N+1) - \log N\end{aligned}$$

We want to separate out the terms dependent on n_j, so we separate out the first term as:

$$\begin{aligned}&\sum_{i \backslash j}^{K} n_i \left(\frac{1}{N} - \frac{1}{N+1}\right) \log n_i \\ &= \sum_{i}^{K} \frac{n_i \log n_i}{N(N+1)} \\ &\quad - \frac{n_j \log n_j}{N(N+1)}\end{aligned}$$

The difference therefore reduces to the following terms, with E representing the terms independent of n_j:

$$\mathcal{H}_j - \mathcal{H} = E + \frac{n_j}{N} \log n_j - \frac{n_j+1}{N+1} \log(n_j+1) - \frac{n_j}{N(N+1)} \log n_j$$

Here, the term independent of n_j, denoted E, is given by:

$$E = \sum_{i}^{K} \frac{n_i \log n_i}{N(N+1)} + \log(N+1) - \log N$$

We further simplify the terms dependent on n_j:

$$\begin{aligned}&\frac{n_j}{N} \log n_j - \frac{n_j+1}{N+1} \log(n_j+1) - \frac{n_j}{N(N+1)} \log n_j \\ &= \frac{N n_j \log n_j + n_j \log n_j - N n_j \log(n_j+1) - N \log(n_j+1) - n_j \log n_j}{N(N+1)}\end{aligned}$$

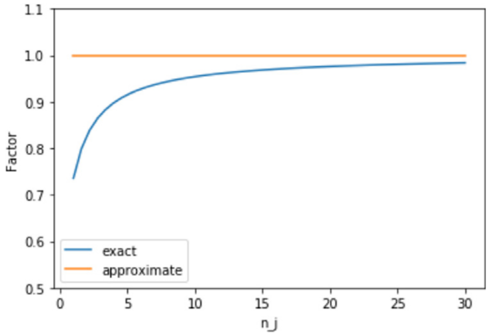


Fig. C.5. Evaluating the approximation: We plot the terms in the exponents of Eq. (C.1) as a function of increasing number of objects in a cluster (n_j). We see that the difference between the approximation and exact value reduces quickly.

$$= -\frac{1}{N(N+1)}\left[Nn_j\log\left(1+\frac{1}{n_j}\right) - N\log(n_j+1)\right]$$

$$= -\frac{n_j\log\left(1+\frac{1}{n_j}\right)}{N+1} - \frac{\log(n_j+1)}{N+1}$$

Taylor expanding the first term to leading order gives:

$$-\frac{n_j\log\left(1+\frac{1}{n_j}\right)}{N+1} \approx -\frac{n_j\left(\frac{1}{n_j}-\frac{1}{2n_j^2}\right)}{N+1}$$

In the large N limit, and correspondingly large n_j limit, we assume $N \approx N+1$ and ignore non-leading terms, to get:

$$-\frac{n_j\left(\frac{1}{n_j}-\frac{1}{2n_j^2}\right)}{N+1} = \frac{-1}{N+1} + \frac{1}{2n_j(N+1)} \approx -\frac{1}{N} \tag{B.1}$$

Appendix C. Evaluating the large N approximation

Here we revisit the approximation made to arrive at Eq. (7) or Eq. (B.1). If we had not made the approximation required to eliminate the extra term, we would have an additional dependence on n_j as follows:

$$\begin{aligned}P(\pi_j)|\pi &\propto \exp\left(-N(\mathcal{H}_j - \mathcal{H})\right)\\ &= \exp(-NE) \times n_j \times \exp(n_j\log(1+1/n_j))\end{aligned}$$

In our simplification, we are making the following approximation:

$$\exp(n_j\log(1+1/n_j)) \sim \exp(1) \tag{C.1}$$

We plot these exponents in Fig. C.5 to give a sense for when this is a good approximation. We see that even at small n_j, the values are relatively close, with the approximation converging quickly.

We restrict our analyses to the correspondence of the conditional distributions $P(\pi_N|\pi_{N-1})$ between the entropy-based distribution and the CRP, rather than directly examining the joint distribution $P(\pi_N)$. This is because computing the normalization factor for the conditional distribution for the entropy-based distribution (*before* making the approximation above) depends on the distribution of objects in the previous step—unlike after we make the approximation when the normalization factor goes to $Ne + 1$. This makes the pre-approximation joint distribution difficult to compute.

References

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the internet. *Glottometrics*, *3*(1), 143–150.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de saint-flour XIII—1983, Vol. 1117* (pp. 1–198). Springer.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions.. *Psychological Review*, *125*(6), 985–1001.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience.. *Psychological Bulletin*, *138*(3), 389.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization.. *Psychological Review*, *103*(3), 566.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Duvenaud, D., Maclaurin, D., & Adams, R. (2016). Early stopping as nonparametric variational inference. In *Artificial intelligence and statistics* (pp. 1070–1077). PMLR.

Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(5), 330–340.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning* (pp. 1050–1059). PMLR.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197.

Gershman, S. J., & Cikara, M. (2021). Structure learning principles of stereotype change. http://dx.doi.org/10.31234/Osf.Io/52f9c, PsyArXiv.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Gottwald, S., & Braun, D. A. (2019). Bounded rational decision-making from elementary computations that reduce uncertainty. *Entropy*, *21*(4), 375.

Griffiths, T. L., & Austerweil, J. (2008). Analyzing human feature learning as nonparametric Bayesian inference. *Advances in Neural Information Processing Systems*, *21*, 97–104.

Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the annual conference of the cognitive science society*.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Sciences*, *7*(2), 217–229.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater, & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 303–328). Oxford, UK: Oxford University Press.

Hennig, P., Osborne, M. A., & Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *471*(2179), Article 20150142.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics*. Cambridge University Press.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Krogh, A., & Hertz, J. (1991). A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, *4*.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung C*, *36*(9–10), 910–912.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, 1–85.

Mandelbrot, B. (1960). The Pareto-Levy law and the distribution of income. *International Economic Review*, *1*(2), 79–106.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

Neal, R. M. (1998). *Markov chain sampling methods for Dirichlet process mixture models: Technical report 9815*, Department of Statistics, University of Toronto.

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network. Computation in Neural Systems*, *7*(2), 333–339.

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K. E., & Tishby, N. (2015). Information-theoretic bounded rationality. arXiv preprint arXiv:1512.06789.

Rosen, K. T., & Resnick, M. (1980). The size distribution of cities: an examination of the Pareto law and primacy. *Journal of Urban Economics*, *8*(2), 165–186.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell Labs Technical Journal*, *27*(3), 379–423.

Shore, J., & Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, *26*(1), 26–37.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, *106*(28), 11478–11483.

Zenon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, *123*, 5–18.