Optimal compression in human concept learning

Nathaniel Imel[1,2] **(nimel@uci.edu)** **Noga Zaslavsky**[1,2] **(nogaz@uci.edu)**

[1]Department of Language Science, University of California, Irvine
[2]Department of Psychology, New York University

Abstract

The computational principles that underlie human concept learning have been debated in the literature for decades. Here, we formalize and test a new perspective that is grounded in rate-distortion theory (RDT), the mathematical theory of optimal (lossy) data compression, which has recently been gaining increasing popularity in cognitive science. More specifically, we characterize optimal conceptual systems as solutions to a special type of RDT problem, show how these optimal systems can generalize to unseen examples, and test their predictions for human behavior in three foundational concept-learning experiments. We find converging evidence that optimal compression may underlie human concept learning. Our work also lends new insight into the relation between learnability and compressibility; integrates prototype, exemplar, and Bayesian approaches to human concepts within the RDT framework; and offers a potential theoretical link between concept learning and other cognitive functions that have been successfully characterized by efficient compression.

Keywords: concept learning; categories; information theory; lossy compression

Introduction

Concepts allow humans to understand, categorize, and navigate a complex world. How people acquire concepts has been a topic of intense study for psychology, leading to a number of influential accounts of category learning emerging over the decades (e.g., Anderson, 1991; Goodman et al., 2008; Kruschke, 1992; Love et al., 2004; Medin & Schaffer, 1978; Rosch, 1973; Shepard et al., 1961; Smith & Medin, 1981). Many of these models have been aimed at explaining how people can learn and deploy concepts flexibly, under finite time and memory. Most relevant to our work is a class of models that formulates concept learning in terms of probabilistic clustering (Anderson, 1991), mapping observations into a set of compressed mental representations which can later be used to predict unobserved features. While Anderson (1991) derived this approach from a rational Bayesian account of human cognition, the information-theoretic optimality of human concepts—namely, the degree to which they constitute *optimally* compressed representations of the environment—has been far less explored (though see Martínez, 2024). Meanwhile, rate-distortion theory (RDT; Berger, 1971; Shannon, 1959), the mathematical theory of optimal data compression, has recently been applied successfully to a wide range of cognitive phenomena, including working memory (Jakob & Gershman, 2023), perception (Bates & Jacobs, 2020; Sims, 2018), decision making (Aridor et al., 2023; Lai & Gershman, 2021), music (Jacoby et al., 2015), and language (Zaslavsky et al., 2018, 2021), as well as to biological information processing more generally (Tkačik & Bialek, 2016). Here, we hypothesize that the same compression theory applies to concept learning. Specifically, we ask: is human concept learning guided by optimal data compression?

Intuitively, our model predicts that concepts are formed by optimizing a tradeoff between maximally compressing observations that an agent had experienced (i.e., training data) and representing the environment as accurately as possible. Here, we consider representation accuracy to consist of two components: reconstructability of observed features, such as the shape and color of objects in the environment, and predictability of a target feature that may not be observed in the future, such as a category's label.

Our general approach shares ideas with multiple existing accounts in the literature, but departs from them in important ways. For example, it is related to exemplar and prototype-based theories in predicting that concepts are represented as centroids in a multidimensional psychological space (e.g., Medin & Schaffer, 1978; Nosofsky, 1986; Reed, 1972; Rosch, 1973). It is also related to rational analysis approaches (e.g., Anderson, 1991; Goodman et al., 2008; Tenenbaum, 1998) in predicting Bayesian behavior under external environmental constraints. Similar to Anderson's model, our model also unifies prototype, exemplar, and Bayesian inference. However, our approach departs from prior work in three main aspects. First, rather than assuming a probabilistic Bayesian model, which requires strong assumptions about the characteristics of the prior and posterior distributions, we derive the full probabilistic structure of a conceptual system as an optimal solution to a RDT problem. The learning mechanism that emerges from this approach is an optimization algorithm for RDT. Second, our approach commits to a specific cost function—the mutual information between the concepts and training examples—and directly minimizes it within the overall objective function we propose. This informational cost is inherent to RDT and has been derived from first principles (Shannon, 1959). It captures representational complexity as the number of bits required for representing the observed samples using the conceptual system. Third, we explain how people may use different strategies to generalize from compressed concepts. In parallel to our work, Martínez (2024)

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

has also proposed an application of RDT to concept learning, and showed how this perspective can integrate prototype and exemplar theories. Our approach differs from Martínez (2024) in three key aspects: (1) we derive a novel RDT optimization problem that generalizes the Information Bottleneck (IB) principle (Tishby et al., 1999), with an analytical characterization of both the optimal prototype and predictive distribution that is associated with each concept; (2) we derive formal connections not only to prototype and exemplar models but also to Bayesian categorization models; and (3) we show how our RDT model can generalize to novel stimuli.

As a first step in testing our approach, we focus on the batch-learning setting, in which a set of training examples is first presented to the learner, and based on that our model predicts a range of optimal conceptual systems with varying complexity-accuracy tradeoffs. This setting is particularly important not only because it corresponds to a prominent experimental paradigm in the concept learning literature, but also because it provides a theoretical bound on the efficiency of conceptual systems that are learned incrementally. While it is possible to extend our model to settings where concepts are adapted incrementally, we leave this important extension to future work. We test our model on human data from three foundational experiments of concept learning with (1) binary features (Medin & Schaffer, 1978; Nosofsky, Palmeri, & McKinley, 1994), (2) continuous features (Nosofsky & Palmeri, 1998), and (3) and varying task difficulty (Nosofsky, Gluck, et al., 1994; Shepard et al., 1961). Our results suggest that optimally compressed concepts may underlie human behavior and offer a path toward further testing our RDT approach in more complex concept learning settings.

Concept learning via optimal lossy compression

We are interested in exploring whether human concepts may be shaped by pressure to form maximally compressed representations of the world that can still support behavioral goals, such as prediction of a target feature. To this end, we turn to the mathematical framework of rate-distortion theory (RDT; Berger, 1971; Shannon, 1959) and derive an application of this framework for concept learning.

Rate-distortion formulation. RDT characterizes the optimal compression schemes that minimize the number of bits that are required for representing a source variable $X \in \mathcal{X}$, drawn from a source distribution $p(x)$, while not exceeding a permissible degree of reconstruction error, or more generally, distortion. A compression scheme is defined by an encoder $q(c|x)$ that maps X to its compressed representation $C \in \mathcal{C}$. In terms of concept learning (Figure 1), we take X to be an observed input stimulus represented in a feature vector space $\mathcal{X} \subseteq \mathbb{R}^n$, and C to be a mentally compressed representation which we consider as a concept. We further assume that each concept is associated with a reconstructed feature vector $\hat{x}_c = g(c)$, given by a decoder $g : \mathcal{C} \to \mathbb{R}^n$. For simplicity, we take $\mathcal{X}$ to be a finite set of observations and $\mathcal{C} = \{1, \ldots, |\mathcal{X}|\}$,

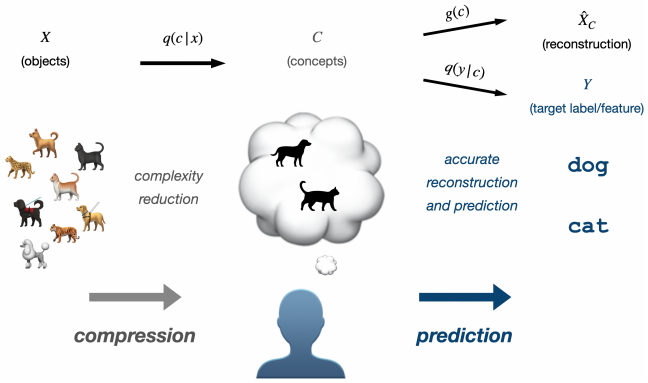

Figure 1: Illustration of our compression model for concept learning (see main text for details).

such that $\hat{x}_c$ could be a prototype or exemplar, depending on the nature of the encoder and decoder.

Given a **distortion** measure $d(x,c)$, RDT characterizes the optimal compression scheme by a stochastic encoder $q(c|x)$ that attains the minimum of

$$I(X;C) + \beta \mathbb{E}[d(X,C)]. \tag{1}$$

The first term is the **complexity** of the conceptual system, defined by the mutual information between the input stimulus and its concept. This quantity is also called the information rate, as it corresponds to the expected number of bits required to mentally represent a stimulus X as concept C. The second term is the expected distortion, taken with respect to $p(x)q(c|x)$. $\beta \geq 0$ is a Lagrange multiplier that specifies the trade-off between minimizing complexity and distortion.

The distortion measure is not directly specified by RDT; different measures will be more or less appropriate, depending on the structure of the domain. Specifying the distortion in our context amounts, in Anderson (1991)'s terms, to making assumptions about the "environment in which cognitive processes evolve." Here, we have assumed that inputs can be represented in a feature vector space, which is a standard assumption in the concept learning literature. In this case, it is natural to consider a quadratic loss between the feature vectors of input stimuli and concepts: $\|x - \hat{x}_c\|^2$. At the same time, achieving goals in the environment may require predicting additional target features that may not always be observed, such as a categorical label $Y \in \{0,1\}$. We therefore assume that each input stimuli induces a distribution over target features, $p(y|x)$, and each concept is also associated with a predictive distribution $q(y|c)$. For any given encoder $q(c|x)$, the ideal predictive distribution is the one that minimizes the cross-entropy loss $l(x,c) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(y|x) \log q(y|c)$. Taking together these two notions of concept fitness, we propose an overall distortion function for concept learning tasks that is a weighted average between the classification loss and the feature loss:

$$d_\alpha(x,c) = \alpha \cdot l(x,c) + (1-\alpha) \cdot \|x - \hat{x}_c\|^2, \tag{2}$$

where $\alpha \in [0,1]$ can be interpreted as an attention parameter that specifies the trade-off between attending to the labels of objects vs. attending to their features.

Optimal conceptual systems. Plugging our distortion function into equation (1) yields a non-standard RDT problem, where $q(y|c)$ and $g(c)$ are being optimized together with the encoder:

$$\min_{q,g} \quad I(X;C) + \beta \mathbb{E}[d_\alpha(X,C)]. \tag{3}$$

When $\alpha = 1$, i.e., when all the attention is shifted to predicting the target feature, this optimization problem becomes the well-known Information Bottleneck (IB) principle (Tishby et al., 1999). In this case, the relevant structure in X is solely determined by the information it maintains on the target feature, or label, without attending to the feature topology. Intuitively, this poses a challenge for generalizing to unseen inputs. When $\alpha = 0$, i.e., when all the attention is shifted to the feature space, this problem is reduced to a rate distortion problem with squared error loss. In this case, however, it is unclear how one may be able to learn task-specific categorization systems for similar inputs, because no information about the task-specific target labels is taken into account. We hypothesize that in between these extremes — of paying attention only to features vs. only to labels — there exist trade-off solutions that may well-describe aspects of human concept learning.

Solving the RDT optimization problem in Eq. (3) gives the following self-consistent equations as a necessary condition for optimality (see Zaslavsky et al., 2021, for a closely related derivation):

$$q(c|x) \propto q(c) \exp\left(-\beta d_\alpha(x,c)\right) \tag{4}$$

$$q(c) = \sum_{x \in \mathcal{X}} p(x) q(c|x) \tag{5}$$

$$q(y|c) = \sum_{x \in \mathcal{X}} p(y|x) q(x|c) \tag{6}$$

$$\hat{x}_c = \sum_x q(x|c) x\,. \tag{7}$$

Eq. (4) characterizes the optimal encoder, which is exponential in the distortion, and Eq. (5) gives the relative weight of each concept. The optimal predictive distribution for each concept, given by Eq. (6), is precisely the Bayesian posterior distribution of Y given C. Finally, Eq. (7) shows that the optimal representative $\hat{x}_c$ of each concept is given by the centroid feature vector. When $\beta = 0$, the optimization in Eq. (3) only minimizes complexity, which amounts to a trivial solution where X and C are independent and $\hat{x}_c = \mathbb{E}_{p(x)}[X]$ for all c, which is the maximally compressed prototypical representation of the input. As β increases there will be more pressure on minimizing distortion and the representation of the inputs will become more refined, including more unique prototypes. As $\beta \to \infty$, the encoder will become deterministic, mapping each input x to its own concept. This extreme amounts to an exemplar representation. Therefore, β allows concepts to continuously evolve from prototypes to exemplars.

Generalization to new stimuli

So far we have showed how to derive optimal conceptual systems for a given sample $\mathcal{X}$ of input stimuli, namely, a training sample. These optimal systems are parameterized by two parameters: the rate-distortion tradeoff β, and the attention parameter α. However, we have not yet shown how these systems can generalize to unseen inputs, as RDT does not address this question. Below, we consider three possible generalization strategies that are based on the assumption that the learned concepts and their associated representations, i.e., equations (5)-(7)), are accessible after learning, but the training encoder (4) may or may not be accessible.

If people have access to the training encoder during evaluation on examples that contain training stimuli, they may simply use it to obtain predictions by marginalizing over concepts: $q(y|x_{\text{train}}) = \sum_c q(c|x_{\text{train}}) q(y|c)$. If they don't have access to the encoder, or need to classify new stimuli that were never seen during training, then one naive approach would be to map x to the nearest concept

$$c_x = \operatorname*{argmin}_c \|x - \hat{x}_c\|\,, \tag{8}$$

and then predict the label based on $q(y|c_x)$. Another approach would be to generalize *softly*. This can be achieved by defining an ad-hoc approximated encoder:

$$\tilde{q}(c|x) \propto q(c) \exp\left(-\beta \|x - \hat{x}_c\|^2\right), \tag{9}$$

which intuitively amounts to shifting all the attention to the observed features by setting $\alpha = 0$. It is important to note that this encoder is not optimal, not even for $\alpha = 0$, and it implicitly represents the previously learned relationship between features and labels through $\hat{x}_c$ and $q(y|c)$. The Bayesian prediction based on this encoder is

$$\tilde{q}(y|x) = \sum_c q(y|c) \cdot \tilde{q}(c|x)\,. \tag{10}$$

These choice points lead us to formulate three hypotheses for generalization strategies:

Strategy 1. Assuming the training encoder is accessible, people can use it to classify previously seen examples. For classifying unseen examples, they use the nearest concept:

$$q_1(y|x) \coloneqq \begin{cases} \sum_c q(c|x) q(y|c) & x \in \mathcal{X} \\ q(y|c_x) & x \notin \mathcal{X} \end{cases} \tag{11}$$

Strategy 2. This strategy is the same as Strategy 1 in assuming that the training encoder is accessible, but for new examples it predicts a softer generalization:

$$q_2(y|x) \coloneqq \begin{cases} \sum_c q(c|x) q(y|c) & x \in \mathcal{X} \\ \tilde{q}(y|x) & x \notin \mathcal{X} \end{cases} \tag{12}$$

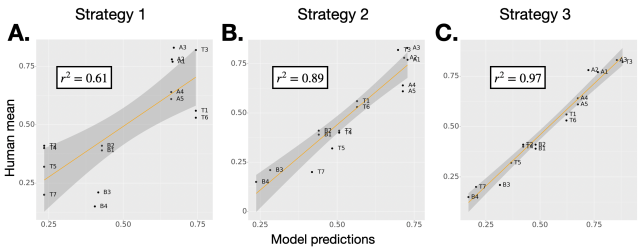

Figure 2: Model predictions vs. human mean responses from NPM94's replication of experiment 2 of MS78. Each point represents a training (A or B) or transfer (T) stimulus. r is the Pearson correlation coefficient.

Strategy 3. Finally, if people do not have access to the training encoder and generalize softly, then their behavior should be captured by.

$$q_3(y|x) \coloneqq \tilde{q}(y|x) \tag{13}$$

Next, we test the predictions of these strategies on human data from three foundational concept learning experiments.

Evaluation on human concept-learning data

To explore the preliminary fit of our model to human data, we consider three classic studies from the concept-learning literature that are often used as initial tests for categorization models. In the experiments we considered, participants were asked to distinguish between two mutually exclusive categories, e.g. A and B. In the first two experiments, there was a distinction between training and transfer stages: during training blocks, participants were presented with each training example once per block and given feedback on their responses, until the dataset could be correctly classified above a threshold. Participants were then asked some time later to classify examples from a larger 'transfer' set, which contained both the original, labeled training examples and new, unlabeled examples that were not seen before.

In the first two experiments we consider, there are per-stimulus, mean human responses available which we use to directly evaluate our model. Recall that the model has only two free parameters, the feature/label attention α and compression tradeoff β. We fit these parameters to the data by maximum likelihood estimation (MLE) for each of the three strategies described above. More specifically, we grid-search 100 values of $\alpha \in [0, 1]$ and roughly 1100 values of $\beta \in [0, 1^{10}]$, selecting for each strategy the pair α, β that maximizes the likelihood of the mean human response. In the last experiment we consider, we do not have access to the per-stimulus human responses. Instead, we rely on a qualitative comparison of our model to the ordinal difficulty of each of the six tasks as reflected in human learning curves.

Binary features. The classical experiments of Medin and Schaffer (1978) (henceforth MS78) are often used as a first

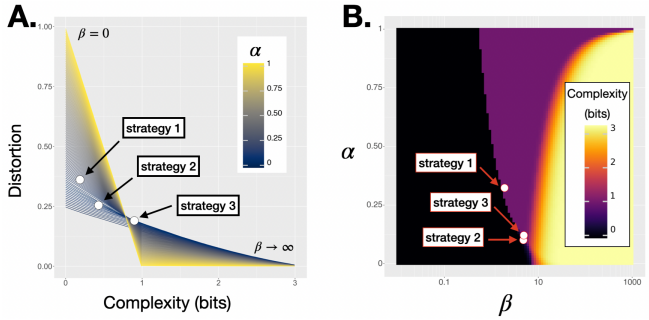

Figure 3: **A**. Compression bounds for the MS78 training stimuli. Colors correspond to the value of α, and for each α, varying β spans a rate-distortion curve. **B**. A heatmap visualizing the complexity of the optimal conceptual systems as a function of α and β. White circles correspond to the best-fitting models for each strategy. Interestingly, the best-fitting models lie on the border of minimal complexity.

test of a classification model's ability to correctly predict prototype enhancement and typicality effects (Posner & Keele, 1968; Reed, 1972). Following Goodman et al. (2008), we consider their second experiment, and human data from the feature-balanced replication by Nosofsky, Palmeri, and McKinley (1994) (henceforth NPM94). The stimuli were drawings of rocket ships varying in four binary-valued dimensions: shape of tail, wings, nose and porthole. The category structure was such that the modal prototype of Category A, encoded as $(0, 0, 0, 0)$, was in the transfer set, labeled T3. The modal prototype of Category B, encoded as $(1, 1, 1, 1)$ was in the training set, labeled B4.

Figure 2 depicts our results. Under Strategy 3, the best-fitting model achieves very strong positive correlation with the observed human responses (Figure 2C, Pearson's $r^2 = 0.97$), and explains 93% of the variance in human data. Under Strategy 2, the best-fitting model does not perform as well, although it still captures much of the structure in the data as seen by the high correlation in Figure 2B. Strategy 1, on the other hand, performs rather poorly. Inspection of Figure 2 reveals that Strategy 1 does not predict prototype enhancement: it is not 'certain enough' about stimuli B4 (the B prototype) and B3, or about T3 (the A prototype) and other A stimuli. However, Strategies 2 and 3 correctly predict these effects. This suggests that people employ a soft, rather than hard, generalization strategy within this framework of concept learning via optimal compression. While our model achieves comparable performance to prior models (e.g., Medin & Schaffer, 1978; Nosofsky, Palmeri, & McKinley, 1994), with slightly lower explained variance, it achieves so with less inductive bias and fewer free parameters. Furthermore, our optimal compression approach provides a theoretical bound that can be used to study the information-theoretic efficiency of *any* conceptual system.

To see this, Figure 3A shows the theoretical bounds derived from our compression model by computing the optimal solu-

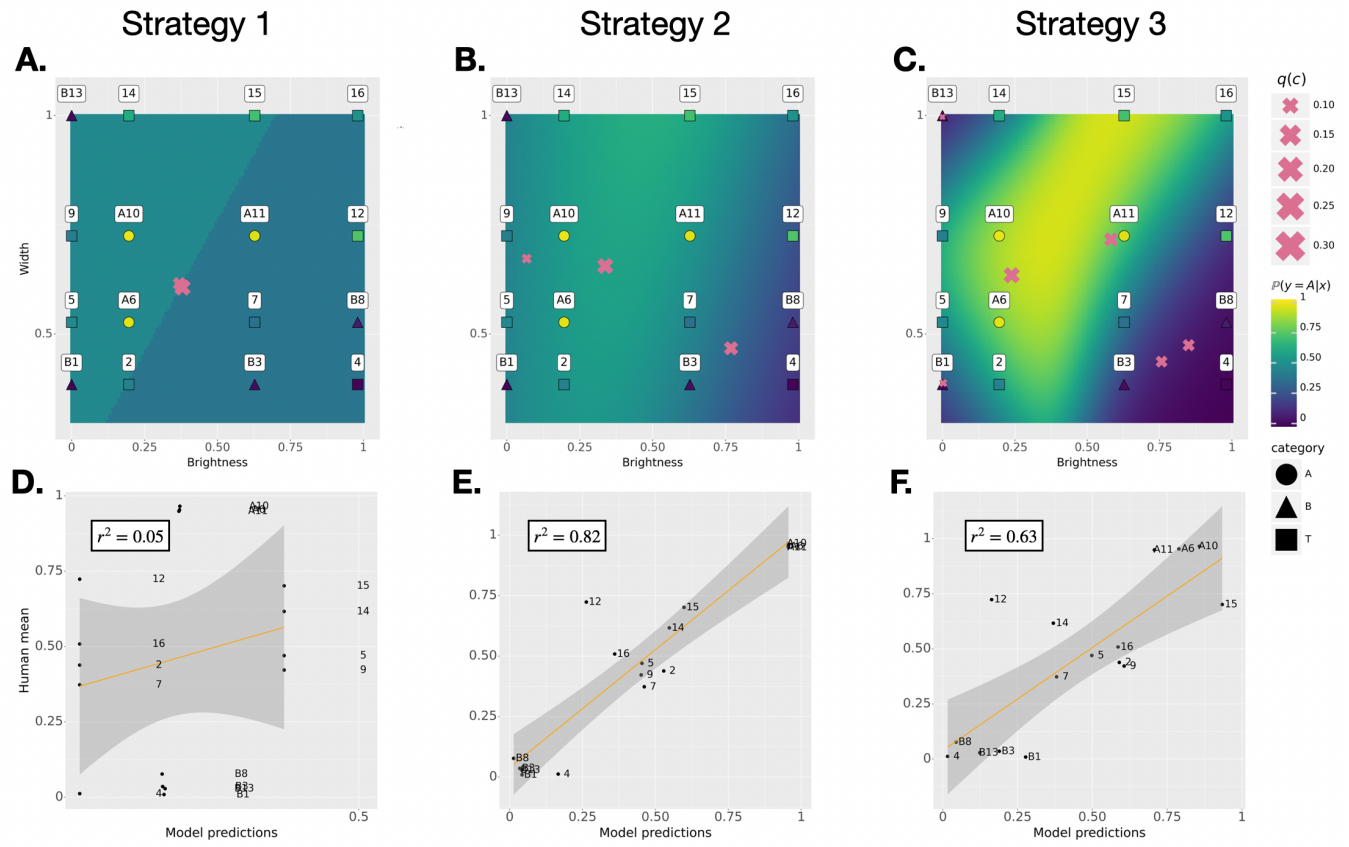

Figure 4: Model predictions and quantitative fits for human predictions in the final experiment block of NP98 (see Movie 1 for an illustration of the evolution of concepts as a function of β, and Movie 2 as a function of α). **A-C**. The 16 stimuli in their 2D feature space (brightness vs. width ratio of ellipses), colored by mean human probability of classifying as Category A, with shape corresponding to their label (A, B, or T for unlabeled/transfer). Background color: the probability of classifying an input as Category A as predicted by each strategy. Model centroids (concepts) are shown as pink Xs, with size corresponding to their weight $q(c)$. **D-F**. Model predictions vs human mean responses as in Figure 2. r is the Pearson correlation coefficient.

tion for densely sampled values of α and β, together with the location of the best-fitting models along these bounds. Here, every colored line is a theoretical bound on efficient compression, where the trade-off between complexity and distortion is specified by β. The color of each line depicts α, which specifies the trade-off between attending to features vs. labels. Importantly, Figure 3A shows that for all three strategies, the best-fitting models are obtained at values of α and β that are far from their extremes, suggesting that attention to both features and labels, as well as active constraints on both compressibility and predictability, may be crucial factors that shape human concept learning. The best model overall (Strategy 1) corresponds to $\beta = 4.85$ and $\alpha = 0.12$, suggesting that more attention may be allocated to the feature space than to the target label. Finally, to get deeper insight into the landscape of optimal solutions, Figure 3B shows how complexity changes as a function of α and β, together with the three strategies in this space. Interestingly, while every single point in this region achieves optimal trade-offs, the best-fitting models lie on the border region between minimal and moderate-to-high complexity.

Continuous features. In contrast to many experiments which modeled category learning in terms of bundles of binary features, Nosofsky and Palmeri (1998) (henceforth NP98) studied how people learn categories with continuous-valued features. In their experiments, participants were presented with ellipses varying in width ratio and brightness. We follow Goodman et al. (2008) in averaging the results of the two experiments to counterbalance the data.[1] The category structure for this task is shown in Figure 4, where examples are colored by the mean human probability of classifying the object as A. Our fit to these data is also encouraging. In this domain, the best strategy appears to be Strategy 2, which achieves strong positive correlation with human responses (Figure 4E, Pearson's $r^2 = .82$) and can explain 82% of the variance in the data. While it may seem at first as if this model

[1] Furthermore, for numerical stability, we renormalize the each of the feature values to lie within $[0, 1]$, though our model is theoretically well-defined for arbitrary real-valued features.

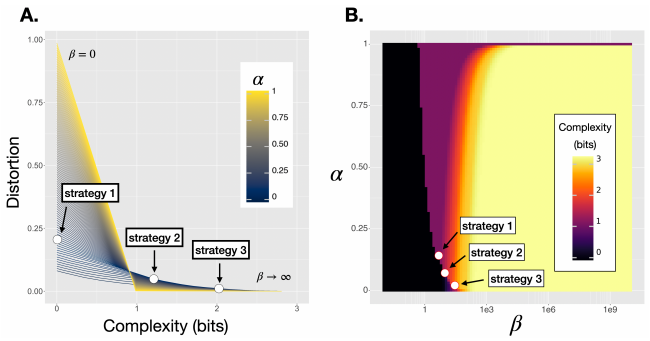

Figure 5: Same as Figure 3 but for the NP98 dataset. As in Figure 3B, human data best corresponds to solutions that lie on the border of minimal complexity.

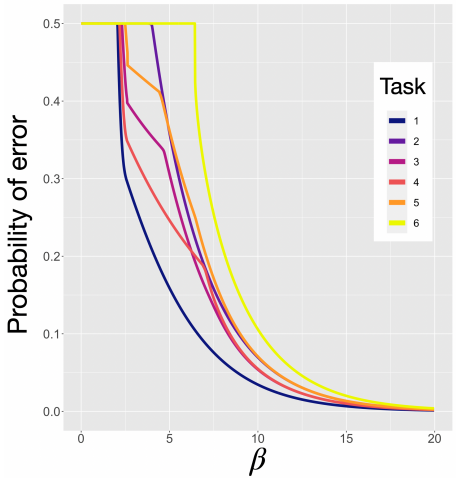

Figure 6: Simulated learning curves for six concept learning tasks from Shepard et al. (1961) of increasing difficulty. Curves show the mean probability of error on the dataset (y-axis) for each value of β (x-axis). These curves show the incompressibility of each of the six tasks' respective concepts.

falls short compared to previous influential models that were able to account for these data almost perfectly, when taking into account the number of free parameters in our model, our results reflect substantial improvement. For example, both the RULEX (NP98) and GCM (Nosofsky, 1986) models both can explain $> 99\%$ of variance in human data, but their AIC values are 86.6 and 109.28, respectively. By comparison, the AIC of Strategies 2 and 3 is 18.58 and 21.17, respectively.

Of equal interest are the qualitative predictions of the model across strategies and across trade-off parameters. Figure 5 shows the theoretical bounds for this domain, similar to Figure 3. As before, we find that the best model is attained at intermediate values of α and β (Figure 5A, Strategy 2), and that all strategies tend to lie near the boundary of minimal complexity (Figure 5B). To get a better sense of how the optimal conceptual systems evolve as a function of β and α, see Movie 1 and Movie 2 respectively. For example, when β is low, the model predicts two distinct values of $\hat{x}_c$ to bifurcate the data. As β increases, so does the number of concepts, which yields lower classification error but also greater complexity. As $\beta \to \infty$, the model predicts centroids that coincide perfectly with training examples.

Predicting task difficulty. Finally, another set of influential results from the literature come from the experiments of Shepard et al. (1961), who studied the relative difficulty of learning for six different categories, constructed from the same basic stimuli. Nosofsky, Gluck, et al. (1994) replicated these experiments and plotted human learning curves. Although our model is not explicitly adaptive, we can simulate learning curves by annealing the trade-off parameter β (i.e., from low values to high) for each task, and inspecting whether the curves qualitatively match the human curves. For space, we show results for $\alpha = 0.25$ and Strategy 3, but our qualitative findings were the same across all other configurations.

Figure 6 shows that ths approach largely recapitulates the ordinal trends of human performance over blocks in Shepard et al. (1961) and Nosofsky, Gluck, et al. (1994), though like many prior models in the literature, Task 2 fails to be predicted to be easier than Tasks 3,4, and 5. This preliminary exploration is interesting, however, as it provides a principled way of analyzing category learning difficulty in terms of lossy incompressibility. That is, for a given probability of error, we can give a lower bound on representational complexity. Complexity increases as learning progress or task difficulty increase. We leave a full analysis (engaging with, e.g., Shepard et al. (1961), p. 42), for future work.

Conclusion

In this work, we have provided a novel approach to human concept learning which is derived from rate distortion theory (RDT), the mathematical theory of lossy data compression. We showed that with minimal assumptions about the representation of different domains, and only two intuitively interpretable free parameters — a feature vs. label attention parameter α, and a complexity-predictability tradeoff parameter β — the model can account for human behavior across three highly influential concept learning experiments. In addition, our theoretical approach provides a new set of tools for characterizing and studying the efficiency of conceptual systems, as well as their evolutionary trajectories as the attention α and tradeoff β vary over time. An important direction that we intend to explore in future work is extending our models to adaptive settings, where concepts are formed as new examples are observed sequentially, as well as applying it to larger-scale domains with more naturalistic inputs. Finally, our work suggests a new theoretical link between models of concept learning and other cognitive functions, such as perception, decision making, and language, to which rate-distortion theory has been successfully applied.

References

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

Aridor, G., Azeredo da Silveira, R., & Woodford, M. (2023). *Information-Constrained Coordination of Economic Behavior *.*
working paper or preprint.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*(5), 891–917.

Berger, T. (1971). *Rate distortion theory; a mathematical basis for data compression.* Prentice-Hall.
Includes bibliographical references (pages 293-301).

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154.

Jacoby, N., Tishby, N., & Tymoczko, D. (2015). An Information Theoretic Approach to Chord Categorization and Functional Harmony. *Journal of New Music Research*, *44*(3), 219–244.

Jakob, A. M., & Gershman, S. J. (2023). Rate-distortion theory of neural coding and its implications for working memory (P. Bays, J. I. Gold, & P. Bays, Eds.). *eLife*, *12*, e79450.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lai, L., & Gershman, S. J. (2021). Chapter Five - Policy compression: An information bottleneck in action selection. In K. D. Federmeier (Ed.), *Psychology of Learning and Motivation* (pp. 195–232, Vol. 74). Academic Press.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332.

Martínez, M. (2024). The Information-Processing Perspective on Categorization. *Cognitive Science*, *48*(2), e13411.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352–369.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(3), 345–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3, Pt.1), 353–363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.

Shannon, C. E. (1959). Coding Theorems for a Discrete Source With a Fidelity Criterion. In *Claude E. Shannon: Collected Papers* (pp. 325–350, Vol. 7). Institute of Radio Engineers, International Convention Record.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Smith, E. E., & Medin, D. L. (1981). 3. The Classical View. In *3. The Classical View* (pp. 22–60). Harvard University Press.

Tenenbaum, J. (1998). Bayesian modeling of human concept learning. *Advances in neural information processing systems*, *11*.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottlneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 368–377.

Tkačik, G., & Bialek, W. (2016). Information Processing in Living Systems. *Annual Review of Condensed Matter Physics*, *7*(1), 89–117.

Zaslavsky, N., Hu, J., & Levy, R. (2021). A Rate–Distortion view of human pragmatic reasoning. *Proceedings of the Society for Computation in Linguistics*, *4*.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.