Optimal Teaching for Limited-Capacity Human Learners

Kaustubh Raosaheb Patil
Affective Brain Lab, UCL
& MIT Sloan Neuroeconomics Lab
kaustubh.patil@gmail.com

Xiaojin Zhu
Department of Computer Sciences
University of Wisconsin-Madison
jerryzhu@cs.wisc.edu

Łukasz Kopeć
Experimental Psychology
University College London
l.kopec.12@ucl.ac.uk

Bradley C. Love
Experimental Psychology
University College London
b.love@ucl.ac.uk

Abstract

Basic decisions, such as judging a person as a friend or foe, involve categorizing novel stimuli. Recent work finds that people's category judgments are guided by a small set of examples that are retrieved from memory at decision time. This limited and stochastic retrieval places limits on human performance for probabilistic classification decisions. In light of this capacity limitation, recent work finds that idealizing training items, such that the saliency of ambiguous cases is reduced, improves human performance on novel test items. One shortcoming of previous work in idealization is that category distributions were idealized in an ad hoc or heuristic fashion. In this contribution, we take a first principles approach to constructing idealized training sets. We apply a machine teaching procedure to a cognitive model that is either limited capacity (as humans are) or unlimited capacity (as most machine learning systems are). As predicted, we find that the machine teacher recommends idealized training sets. We also find that human learners perform best when training recommendations from the machine teacher are based on a limited-capacity model. As predicted, to the extent that the learning model used by the machine teacher conforms to the true nature of human learners, the recommendations of the machine teacher prove effective. Our results provide a normative basis (given capacity constraints) for idealization procedures and offer a novel selection procedure for models of human learning.

1 Introduction

Judging a person as a friend or foe, a mushroom as edible or poisonous, or a sound as an \l\ or \r\ are examples of categorization tasks. Category knowledge is often acquired based on examples that are either provided by a teacher or past experience. One important research challenge is determining the best set of examples to provide a human learner to facilitate learning and use of knowledge when making decisions, such as classifying novel stimuli. Such a teacher would be helpful in a pedagogical setting for curriculum design [1, 2].

Recent work suggests that people's categorization decisions are guided by a small set of examples retrieved at the time of decision [3]. This limited and stochastic retrieval places limits on human performance for probabilistic classification decisions, such as predicting the winner of a sports contest or classifying a mammogram as normal or tumorous [4]. In light of these capacity limits, Giguère and Love [3] determined and empirically verified that humans perform better at test after being

trained on *idealized* category distributions that minimize the saliency of ambiguous cases during training. Unlike machine learning systems that can have unlimited retrieval capacity, people performed better when trained on non-representative samples of category members, which is contrary to common machine learning practices where the aim is to match training and test distributions [5].

One shortcoming of previous work in idealization is that category distributions were idealized in an ad hoc or heuristic fashion, guided only by the intuitions of the experimenters in contrast to a rigorous systematic approach. In this contribution, we take a first principles approach to constructing idealized training sets. We apply a machine teaching procedure [6] to a cognitive model that is either limited capacity (as humans are) or unlimited capacity (as most machine learning systems are). One general prediction is that the machine teacher will idealize training sets. Such a result would establish a conceptual link between idealization manipulations from psychology and optimal teaching procedures from machine learning [7, 6, 8, 2, 9, 10, 11]. A second prediction is that human learners will perform best with training sets recommended by a machine teacher that adopts a limited capacity model of the learner. To the extent that the learning model used by the machine teacher conforms to the true nature of human learners, the recommendations of the machine teacher should prove more effective. This latter prediction advances a novel method to evaluate theories of human learning. Overall, our work aims to provide a normative basis (given capacity constraints) for idealization procedures.

2 Limited- and Infinite-Capacity Models

Although there are many candidate models of human learning (see [12] for a review), to cement the connection with prior work [3] and to facilitate evaluation of model variants differing in capacity limits, we focus on exemplar models of human learning. Exemplar models have proven successful in accounting for human learning performance [13, 14], are consistent with neural representations of acquired categories [15], and share strong theoretical connections with machine learning approaches [16, 17]. Exemplar models represent categories as a collection of experienced training examples. At the time of decision, category examples (i.e., exemplars) are activated (i.e., retrieved) in proportion to their similarity to the stimulus. The category with the greatest total similarity across members tends to be chosen as the category response. Formally, the categorization problem is to estimate the label $\hat{y}$ of a test item x from its similarity with the training exemplars $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Exemplar models are consistent with the notion that people stochastically and selectively sample from memory at the time of decision. For example, in the Exemplar-Based Random Walk (EBRW) model [18], exemplars are retrieved sequentially and stochastically as a function of their similarity to the stimulus. Retrieved exemplars provide evidence for category responses. When accumulated evidence (i.e., retrieved exemplars) for a response exceeds a threshold, the corresponding response is made. The number of steps in the diffusion process is the predicted response time.

One basic feature of EBRW is that not all exemplars in memory need feed into the decision process. As discussed by Giguère and Love [3], finite decision thresholds in EBRW can be interpreted as a capacity limit in memory retrieval. When decision thresholds are finite, a limited number of exemplars are retrieved from memory. When capacity is limited in this fashion, models perform better when training sets are idealized. Idealization reduces the noise injected into the decision process by limited and stochastic sampling of information in memory.

We aim to show that a machine teacher, particularly one using a limited-capacity model of the learner, will idealize training sets. Such a result would provide a normative basis (given capacity constraints) for idealization procedures. To evaluate our predictions, we formally specify a limited- and unlimited-capacity exemplar model. Rather than work with EBRW, we instead choose a simpler mathematical model, the Generalized Context Model (GCM, [14]), which offers numerous advantages for our purposes. As discussed below, a parameter in GCM can be interpreted as specifying capacity and can be related to decision threshold placement in EBRW's drift-diffusion process.

Given a finite training set (or a teaching set, we will use the two terms interchangeably) $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and a test item (i.e., stimulus) x, GCM estimates the label probability as:

$$\hat{p}(y = 1 \mid x, D) \quad = \quad \frac{\left(b + \sum_{i \in D: y_i = 1} e^{-c\, d(x, x_i)}\right)^{\gamma}}{\left(b + \sum_{i \in D: y_i = 1} e^{-c\, d(x, x_i)}\right)^{\gamma} + \left(b + \sum_{i \in D: y_i = -1} e^{-c\, d(x, x_i)}\right)^{\gamma}} \quad (1)$$

where d is the distance function that specifies the distance (e.g., the difference in length between two line stimuli) between the stimulus x and exemplar x_i, c is a scaling parameter that specifies the rate at which similarity decreases with distance (i.e. the bandwidth parameter for a kernel), and the parameter b is background similarity, which is related to irrelevant information activated in memory. Critically, the response scaling parameter, γ, has been shown to bear a relationship to decision threshold placement in EBRW [18]. In particular, Equation 1 is equivalent to EBRW's mean response (averaged over many trials) with decision threshold bounds placed γ units away for the starting point for evidence accumulation. Thus, GCM with a low value of γ can be viewed as a limited capacity model, whereas GCM with a high value for γ converges to the predictions of an infinite capacity model. These two model variations (low and high γ as surrogates for low- and high-capacity) will figure prominently in our study and analyses.

To select a binary response, the learner samples a label according to the probability $\hat{y} \sim \text{Bernoulli}(\hat{p}(y = 1 \mid x, D))$. Therefore, the learner makes stochastic predictions. When measuring the classification error of the learner, we will take expectation over this randomness. Let the distance function be $d(x_i, x_j) = |x_i - x_j|$. Thus a GCM learner can be represented using three parameters $\{b, c, \gamma\}$.

3 Machine Teaching for the GCM Learners

Machine teaching is an inverse problem of machine learning. Given a learner and a test distribution, machine teaching designs a small (typically non-iid) teaching set D such that the learner trained on D has the smallest test error [6]. The machine teaching framework poses an optimization problem:

$$\min_{D \in \mathbb{D}} \ \text{loss}(D) + \text{effort}(D). \tag{2}$$

The optimization is over D, the teaching set that we present to the learner. For our task, $D = (x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in [0, 1]$ represents the 1D feature of the i^{th} stimulus, and $y_i \in \{-1, 1\}$ represents the i^{th} label. The search space $\mathbb{D} = \{(\mathcal{X} \times \mathcal{Y})^n : n \in \mathbb{N}\}$ is the (infinite) set of finite teaching sets. Importantly, D is not required to consist of iid items drawn from the test distribution $p(x, y)$. Rather, D will usually contain specially arranged items. This is a major difference to standard machine learning.

Since we want to minimize classification error on future test items, we define the teaching loss function to be the generalization error:

$$\text{loss}(D) = \mathbb{E}_{(x,y)\sim p(x,y)} \mathbb{E}_{\hat{y}\sim\hat{p}(y|x,D)} \mathbb{1}_{y\neq\hat{y}}. \tag{3}$$

The first expectation is with respect to the test distribution $p(x, y)$. That is, we still assume that test items are drawn iid from the test distribution. The second expectation is w.r.t. the stochastic predictions that the GCM learner makes. Note that the teaching set D enters the loss() function through the GCM model $\hat{p}(y \mid x, D)$ in (1). We observe that:

$$\begin{aligned} \text{loss}(D) &= \mathbb{E}_{x\sim p(x)}\, p(y = 1 \mid x)\hat{p}(y = -1 \mid x, D) + p(y = -1 \mid x)\hat{p}(y = 1 \mid x, D) \\ &= \int \left(\frac{1 - 2p(y = 1 \mid x)}{1 + \left(\frac{b + \sum_{i \in D: y_i = -1} e^{-c\ d(x, x_i)}}{b + \sum_{i \in D: y_i = 1} e^{-c\ d(x, x_i)}} \right)^{\gamma}} + p(y = 1 \mid x) \right) p(x) dx. \end{aligned} \tag{4}$$

The teaching effort function effort(D) is a powerful way to specify certain preferences on the teaching set space $\mathbb{D}$. For example, if we use effort(D) $= |D|$ the size of D then the machine teaching problem (2) will prefer smaller teaching sets. In this paper, we use a simple definition of effort(): effort(D) $= 0$ if $|D| = n$, and ∞ otherwise. This infinity indicator function simply acts as a hard constraint so that D must have exactly n items. Equivalently, we may drop this effort() term from (2) altogether while requiring the search space $\mathbb{D}$ to consist of teaching sets of size exactly n.

In this paper, we consider test distributions $p(x, y)$ whose marginal on x has a special form. Specifically, we assume that $p(x)$ is a uniform distribution over m distinct test stimuli $z_1, \ldots, z_m \in [0, 1]$. In other words, there are only m distinct test stimuli. The test label y for stimuli z_j in any given test set is randomly sampled from $p(y \mid z_j)$. Besides matching the actual behavioral experiments,

this discrete marginal test distribution affords a further simplification to our teaching problem: the integral in (4) is replaced with summation:

$$\min_{x_1 \ldots x_n \in [0,1]; y_1 \ldots y_n \in \{-1,1\}} \frac{1}{m} \sum_{j=1}^{m} \left(\frac{1 - 2p(y = 1 \mid z_j)}{1 + \left(\frac{b + \sum_{i:y_i=-1} e^{-c\, d(z_j, x_i)}}{b + \sum_{i:y_i=1} e^{-c\, d(z_j, x_i)}} \right)^{\gamma}} + p(y = 1 \mid z_j) \right). \tag{5}$$

It is useful to keep in mind that $y_1 \ldots y_n$ are the training item labels that we can design, while y is a dummy variable for the stochastic test label.

In fact, equation (5) is a mixed integer program because we design both the continuous training stimuli $x_1 \ldots x_n$ and the discrete training labels $y_1 \ldots y_n$. It is computationally challenging. We will relax this problem to arrive at our final optimization problem. We consider a smaller search space $\mathbb{D}$ where each training item label y_i is uniquely determined by the position of x_i w.r.t. the true decision boundary $\theta^* = 0.5$. That is, $y_i = 1$ if $x_i \geq \theta^*$ and $y_i = -1$ if $x_i < \theta^*$. We do not have evidence that this reduced freedom in training labels adversely affect the power of the teaching set solution. We now removed the difficult discrete optimization aspect, and arrive at the following continuous optimization problem to find an optimal teaching set (note the changes to selector variables i):

$$\min_{x_1 \ldots x_n \in [0,1]} \frac{1}{m} \sum_{j=1}^{m} \left(\frac{1 - 2p(y = 1 \mid z_j)}{1 + \left(\frac{b + \sum_{i:x_i<0.5} e^{-c\, d(z_j, x_i)}}{b + \sum_{i:x_i \geq 0.5} e^{-c\, d(z_j, x_i)}} \right)^{\gamma}} + p(y = 1 \mid z_j) \right). \tag{6}$$

4 Experiments

Using the machine teacher, we derive a variety of optimal training sets for low- and high-capacity GCM learners. We then evaluate how humans perform when trained on these recommended items (i.e. training sets). The main predictions are that the machine teacher will idealize training sets and that humans will perform better on optimal training sets calculated using the low-capacity GCM variant. In what follows, we first specify parameter values for the GCM variants, present the optimal teaching sets we calculate, and then discuss human experiments.

4.1 Specifying GCM parameters

The machine teacher requires a full specification of the learner, including its parameters. Parameters were set for the low-capacity GCM model by fitting the behavioral data from Experiment 2 of Giguère and Love [3]. GCM was fit to the aggregated data representing an average human learner by solving the following optimization problem:

$$\{\hat{b}, \hat{c}, \hat{\gamma}\} = \arg\min_{\hat{b}, \hat{c}, \hat{\gamma}} \sum_{i \in X^{(1)}} \left(g^{(1)}(x_i) - f^{(1)}(x_i) \right)^2 + \sum_{j \in X^{(2)}} \left(g^{(2)}(x_j) - f^{(2)}(x_j) \right)^2 \tag{7}$$

where $X^{(1)}$ and $X^{(2)}$ are sets of unique test stimuli for the two training conditions (actual and idealized) in Experiment 2. We define two functions to describe the estimated and empirical probabilities, respectively: $g^{(cond)}(x_i) = p(y_i = 1 \mid x_i, D^{(cond)})$, $f^{(cond)}(x_i) = \frac{\sum_{j \in D^{(cond)}: y_j = 1} \mathbb{1}(x_j = x_i)}{\sum_{j' \in D^{(cond)}} \mathbb{1}(x'_j = x_i)}$. The function g above is defined using GCM in Equation 1. We solved Equation 7 to obtain the low-capacity GCM parameters that best capture human performance $\{\hat{b}, \hat{c}, \hat{\gamma}\} = \{5.066, 2.964, 4.798\}$. We define a high-capacity GCM by only changing the $\hat{\gamma}$ parameter, which is set an order of magnitude higher at $\hat{\gamma} = 47.98$.

4.2 Optimal Teaching Sets

The machine teacher was used to generate a variety of training sets that we evaluated on human learners. All training sets had size $n = 20$, which was chosen to maximize expected differences in human test performance across training sets. All conditions involved the same test conditional

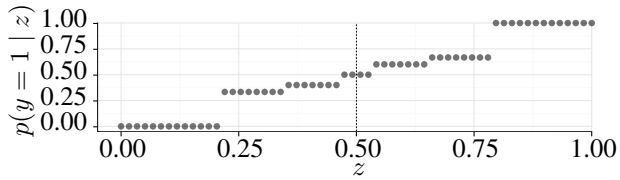

Figure 1: The test conditional distribution. Each point shows a test item z_i and its conditional probability to be in the category $y = 1$. The vertical dashed line shows the location of the true decision boundary $\theta^* = 0.5$.

distribution $p(y \mid x)$ (see Figure 1). The test set consisted of $m = 60$ representative items evenly spaced over the stimulus domain $[0, 1]$ with a probabilistic category structure. The conditional distribution $p(y = 1 \mid x = z_j)$ for $j = 1 \ldots 60$ was adapted from a related study [3]. We then solved the machine teaching problem (6) to obtain the optimal teaching sets for low- and high-capacity learners.

The optimal training set for the low-capacity GCM places items for each category in a clump far from the boundary (see Figure 2 for the optimal training sets). We refer to this optimal training set as *Clump-Far*. The placement of these items far from the boundary reflects the low-capacity (i.e., low γ value) of the GCM. By separating the items from the two categories, the machine teacher makes it less likely that low-capacity GCM will erroneously retrieve items from the opposing category at the time of test. As predicted, the machine teacher idealized the Clump-Far training set.

A mathematical property of the high-capacity GCM suggests that it is sensitive only to the placement of training items adjacent to the decision boundary θ^* (all other training items have exponentially small influence). Therefore, for the high-capacity model up to computer precision, there is no unique optimal teaching set but rather a family of optimal sets (i.e., multiple teaching sets with the same loss or expected test error). We generated two training sets that are both optimal for the high-capacity model. The *Clump-Near* training set has one clump of similar items for each category close to the boundary. In contrast, the *Spread* training set uniformly spaces items outward, mimicking the idealization procedure in Giguère and Love [3]. We also generated *Random* teaching sets by sampling from the joint distribution $U(x)p(y \mid x)$, where $U(x)$ is uniform in $[0, 1]$ and $p(y \mid x)$ is the test conditional distribution. Note Random is the traditional iid training set in machine learning. The test error of the low- and high-capacity GCM under Random teaching sets was estimated by generating 10,000 random teaching sets.

Table 1 shows that Clump-Far outperforms other training sets for the low-capacity GCM. In contrast, Clump-Far, Clump-Near, and Spread are all optimal for high-capacity GCM, reflecting the fact that for high-capacity GCM the symmetry of the inner-most training item pair about the true decision boundary θ^* determines the learned model. Not surprisingly, Random teaching sets lead to suboptimal test errors on both low- and high-capacity GCM.

Table 1: Loss (i.e. test error) for different teaching sets on low- and high-capacity GCM. Note the smallest loss 0.216 matches the optimal Bayes error rate.

GCM Model	Clump-Far	Spread	Clump-Near	Random
Low-capacity	0.245	0.261	0.397	M=0.332, SD=0.040
High-capacity	0.216	0.216	0.216	M=0.262, SD=0.066

In summary, we produced four kinds of teaching sets: (1) Clump-Far which is the optimal teaching set for the low-capacity GCM, (2) Spread, (3) Clump-Near, the three are all optimal teaching sets for the high-capacity GCM, and (4) Random. The next section discusses how human participants fair with each of these four training sets. Consistent with our predictions, the machine teacher's choices idealized the training sets with parallels to the idealization procedures used in Giguère and Love [3]. They found that human learners benefited when within category variance was reduced (akin to clumping in Clump-Far and Clump-Near), training items were shifted away from the category boundary (akin to Clump-Far), and feedback was idealized (as in all the machine teaching sets considered). Their actual condition in which training sets were not idealized resembles the Random condition here. As hoped, low-capacity and high-capacity GCM make radically different

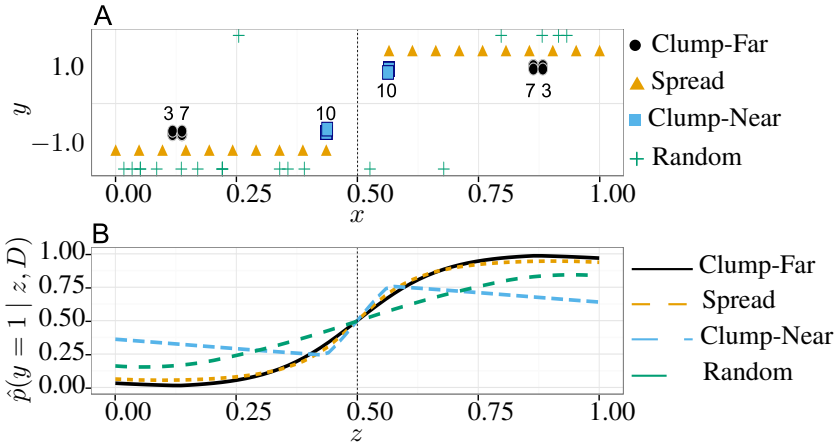

Figure 2: (A) The teaching sets. The points show the machine teaching sets. Overlapping training points are shown as clumps along with the number of items. A particular Random teaching set is shown. All training labels y were in $\{1, -1\}$, but dithered vertically for viewing clarity. (B) The predictive distribution $\hat{p}(y = 1 \mid z, D)$ produced by the low-capacity GCM given a teaching set D. The vertical dashed lines show the position of the true decision boundary θ^*. The curves for the high-capacity GCM were omitted for space.

predictions. Whereas high-capacity GCM is insensitive to variations across the machine teaching sets, low-capacity GCM should perform better under Clump-Far and Spread. The Clump-Near set leads to more errors in low-capacity GCM because items are confusable in memory and therefore limited samples from memory can lead to suboptimal classification decisions. In the next section, we evaluate how humans perform with these four training sets, and compare human performance to that of low- and high-capacity GCM.

4.3 Human Study

Human participants were trained on one of the four training sets: Clump-Far, Spread, Clump-Near, and Random. Participants in all four conditions were tested (no corrective feedback provided) on the $m = 60$ grid test items $z_1 \ldots z_m$ in $[0, 1]$.

Participants. US-based participants ($N = 600$) were recruited via Amazon Mechanical Turk, a paid online crowd-sourcing platform, which is an effective method for recruiting demographically diverse samples [19] and has been shown to yield results consistent with decision making studies in the laboratory [20]. In our sample, 297 of the 600 participants were female and the average age was 34.86. Participants were paid $1.00 for completing the study with the highest performing participant receiving a $20 bonus.

Design. Participants were randomly assigned to one of the four teaching conditions (see Figure 2). Notice that feedback was deterministic in all the teaching sets provided by the machine teacher, but was probabilistic as a function of stimulus for the Random condition. For the Random condition, each participant received a different sample of training items. The test set always consisted of 60 stimuli (see Figure 1). In both training and test trials, stimuli were presented sequentially in a random order (without replacement) determined for each participant.

Materials and Procedure. The stimuli were horizontal lines of various lengths. Participants learned to categorize these stimuli. The teaching sets values $x_i \in [0, 1]$ were converted into pixels by multiplying it by 400 and adding an offset. The offset for each participant was a uniformly selected random number from 30 to 100. As the study was performed online (see below), screen size varied across participants (height $\bar{x}$=879.16, s=143.34 and width $\bar{x}$=1479.6, s=271.04).

During the training phase, on every trial, participants were instructed to fixate on a small cross appearing in a random position on the screen. After 1000 ms, a line stimulus replaced the cross at the same position. Participants were then to indicate their category decision by pressing a key ("F" or "J") as quickly as possible without sacrificing accuracy. Once the participant responded, the stimulus

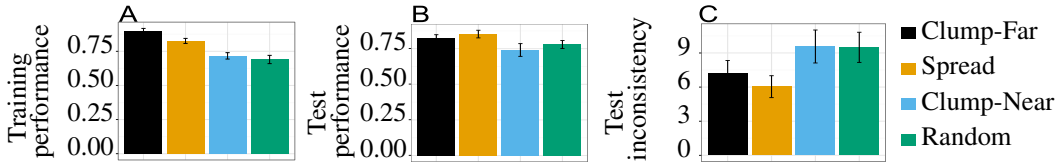

Figure 3: Human experiment results. Each bar corresponds to one of the training conditions. (A) The proportion of agreement between the individual training responses with the Bayes classifier. (B) The proportion of agreement between the individual test responses with the Bayes classifier. (C) Inconsistency in individual test responses. The error bars are 95% confidence intervals.

was immediately replaced by a feedback message ("Correct" or "Wrong"), which was displayed for 2000 ms. The screen coordinates (horizontal/vertical) defining the stimulus (i.e., fixation cross and line) position were randomized on each trial to prevent participants from using marks or smudges on the screen as an aid. Participants completed 20 training trials.

The procedure was identical for test trials, except corrective feedback was not provided. Instead, "Thank You!" was displayed following a response. The test phase consisted of 60 trials. At the end of the test phase each subject was asked to discriminate between the short and long lines from the Clump-Near training set (i.e. $x = 0.435$ and $x = 0.565$, closest stimuli in the deterministically labeled training sets). Both lines were presented side-by-side, with their order counterbalanced between participants. Each participant was asked to indicate which one of those is longer.

Results. It is important that people could perceptually discriminate the categories for the exemplars close to the boundary, especially for the Clump-Near condition in which all the exemplars are close to the boundary. At the end of the main study, this was measured by asking each participant to indicate the longer line between the two. Overall 97% participants correctly indicated the longer line. This did not differ across conditions, $F(3, 596) < 0.84, p \approx 0.47$.

The optimal (i.e. Bayes) classifier deterministically assigns correct class label $\hat{y} = \text{sign}(x - \theta^*)$ to an item x. The agreement between training responses and the optimal classifier were significantly different across the four teaching conditions, $F(3, 596) = 66.97, p < 0.05$. As expected, the random sets resulted in the lowest accuracy (M=65.2%) and the Clump-Far condition resulted in the highest accuracy (M=89.9%) (Figure 3A).

Figure 3B shows how well the test responses agree with the Bayes classifier. The proportional agreement was significantly different across conditions, $F(3, 596) = 9.16, p < 0.05$. The Clump-Far and Spread conditions were significantly different from the Clump-Near condition, $t(228.05) = 3.22, p < 0.05$ and $t(243.84) = 4.21, p < 0.05$, respectively and the Random condition, $t(290.84) = 2.39, p < 0.05$ and $t(297.37) = 3.71, p < 0.05$, respectively. The Clump-Far and the Spread conditions did not differ, $t(294.32) = 1.55, p \approx 0.12$. This result shows that the subjects in the Clump-Far and Spread conditions performed more similar to the Bayes classifier than the subjects in the other two conditions.

Individual test response inconsistency can be calculated using number of neighboring stimuli that are categorized in opposite categories [3]. This measure of inconsistency attempts to quantify the stochastic memory retrieval and higher inconsistency reflects more noisy memory sampling. The inconsistency significantly differed between the conditions, $F(3, 596) = 7.73, p < 0.05$ (Figure 3C). Both Clump-Far and Spread teaching sets showed lower inconsistency, suggesting that those teaching sets lead to less noisy memory sampling. The inconsistencies for these two conditions did not differ significantly, two-sample t test, $t(290.42) = 1.54, p \approx 0.12$. Inconsistencies in conditions Clump-Far and Spread significantly differed from Clump-Near, $t(281.7) = -2.53, p < 0.05$ and $t(291.04) = -2.58, p < 0.05$, respectively and Random, $t(259.18) = -3.98, p < 0.05$ and $t(272.12) = -4.14, p < 0.05$, respectively.

We then calculated *test loss* for each subject as $\sum_{i=1}^{m} (1 - p(h_i \mid z_i))$ where h_i is the response for the stimulus z_i. Figure 4 compares the observed and estimated test performance (i.e. $1 - \text{loss}()$) in four conditions. Overall, human performance is more closely followed by the low-capacity GCM. The human performance across four conditions was significantly different, $F(3, 596) = 11.15, p < 0.05$. The conditions Clump-Far and Spread did not significantly differ, $t(295.96) = -0.8, p \approx$

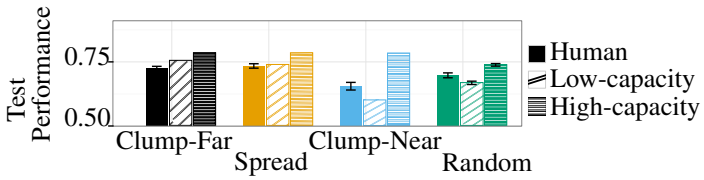

Figure 4: Empirical test performance of human learners for low- and high-capacity GCM on four teaching conditions. Test performance is measured as $1 - \text{loss}()$ (see (3)). Humans follow the low-capacity GCM more closely. The error bars are 95% confidence intervals.

0.42. Test performance in conditions Clump-Far and Spread significantly differed from Clump-Near condition, $t(226.9) = 4.12, p < 0.05$ and $t(287.97) = 2.19, p < 0.05$, respectively and Random condition, $t(238.41) = 4.59, p < 0.05$ and $t(294.72) = 2.85, p < 0.05$, respectively. Humans performed significantly worse in the Clump-Near condition than in the Random condition, $t(253.94) = -2.394, p < 0.05$. A similar pattern was observed for the low-capacity GCM while the opposite for the high-capacity GCM. Inconsistency, as defined above, significantly correlated with the test loss, Pearson's $r = 0.56, t(148) = 8.34, p < 0.05$. Taken together, these results provide support for the low-capacity account of human decision making [3].

In order to check whether the variability within the training set is predictive of test performance we correlated the observed test loss with the estimated loss for the subjects in the Random condition. We observed a significant correlation between the test loss and the estimated loss for both low- and high-capacity models, Pearson's $r = 0.273, t(148) = 3.45, p < 0.05$ and $r = 0.203, t(148) = 2.52, p < 0.05$, respectively. This result points out that due to their limited capacity human learners benefit from lower variability in the training sets, i.e. idealization.

The individual median reaction time in the training phase significantly differed across teaching conditions, $F(3, 596) = 10.66, p < 0.05$. The training median reaction time for the Clump-Far condition was the shortest (M=761 ms, SD=223) and differed significantly from all other conditions, two-sample t tests, all $p < 0.05$. Other conditions did not differ significantly from each other. The individual median reaction times in the test phase (M=767 ms, SD=187) did not differ across teaching conditions, $F(3, 596) = 0.95, p \approx 0.42$.

Taken together, our results suggest that the recommendations of the machine teacher for the low-capacity GCM are indeed effective for human learners. Furthermore, the observed lower inconsistency in this condition suggests that machine teacher is performing idealization which aids by reducing noise in the stochastic memory sampling process.

5 Discussion

A major aim of cognitive science is to understand human learning and to improve learning performance. We devised an optimal teacher for human category learning, a fundamental problem in cognitive science. Based on recent research we focused on GCM which models limited human capacity of exemplar retrieval during decision making. We developed the optimal teaching sets for the low- and high-capacity variants of the GCM learner. By using a 1D category learning task, we have shown that the optimal teaching set for the low-capacity GCM is clumped, symmetrical and located far from the decision boundary, which is intuitively easy to learn. This provides a normative basis (given capacity limits) for the idealization procedures that reduce saliency of ambiguous cases [2, 3]. The optimal teaching set indeed proved effective for human learning.

Future work will pursue several extensions. One interesting topic not considered here is how the order of training examples affects learning. One possibility is that the optimal teacher will recommend easy examples earlier in training and then gradually progress to harder cases [2, 21]. Another important extension is use of multi-dimensional stimuli.

Acknowledgments

The authors are thankful to the anonymous reviewers for their comments. This work is partly supported by the Leverhulme Trust grant RPG-2014-075 to BCL, National Science Foundation grant IIS-0953219 to XZ and WT-MIT fellowship 103811AIA to KRP.

References

[1] P Shafto and N Goodman. A Bayesian Model of Pedagogical Reasoning. In *AAAI Fall Symposium: Naturally-Inspired Artificial Intelligence'08*, pages 101–102, 2008.

[2] Y Bengio, J Louradour, R Collobert, and J Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, New York, USA, June 2009. ACM Press.

[3] G Giguère and B C Love. Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, 110(19):7613–8, May 2013.

[4] A N Hornsby and B C Love. Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, 3:72–76, 2014.

[5] J Q Candela, M Sugiyama, A Schwaighofer, and N D Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, first edit edition, 2009.

[6] X Zhu. Machine Teaching for Bayesian Learners in the Exponential Family. In *Advances in Neural Information Processing Systems*, pages 1905–1913, 2013.

[7] S A Goldman and M J Kearns. On the Complexity of Teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

[8] F Khan, X Zhu, and B Mutlu. How Do Humans Teach: On Curriculum Learning and Teaching Dimension. In *Advances in Neural Information Processing Systems*, pages 1449–1457, 2011.

[9] F J Balbach and T Zeugmann. Recent Developments in Algorithmic Teaching. In A H Dediu, A M Ionescu, and C Martín-Vide, editors, *Language and Automata Theory and Applications*, volume 5457 of *Lecture Notes in Computer Science*, pages 1–18. Springer, Berlin-Heidelberg, March 2009.

[10] M Cakmak and M Lopes. Algorithmic and Human Teaching of Sequential Decision Tasks. In *AAAI Conference on Artificial Intelligence (AAAI-12)*, July 2012.

[11] R Lindsey, M Mozer, W J Huggins, and H Pashler. Optimizing Instructional Policies. In *Advances in Neural Information Processing Systems*, pages 2778–2786, 2013.

[12] B C Love. Categorization. In K N Ochsner and S M Kosslyn, editors, *Oxford Handbook of Cognitive Neuroscience*, pages 342–358. Oxford University Press, 2013.

[13] D L Medin and M M Schaffer. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.

[14] R M Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General*, 115(1):39–61, March 1986.

[15] M L Mack, A R Preston, and B C Love. Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, 23:2023–2027, 2013.

[16] Y Chen, E K Garcia, M R Gupta, A Rahimi, and L Cazzanti. Similarity-based Classification: Concepts and Algorithms. *The Journal of Machine Learning Research*, 10:747–776, December 2009.

[17] F Jakel, B Scholkopf, and F A Wichmann. Does cognitive science need kernels? *Trends in Cognitive Science*, 13(9):381–388, 2009.

[18] R M Nosofsky and T J Palmeri. An exemplar-based random walk model of speeded classification. *Psychological review*, 104(2):266–300, April 1997.

[19] M Buhrmester, T Kwang, and S D Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, February 2011.

[20] M J C Crump, J V McDonnell, and T M Gureckis. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, January 2013.

[21] H Pashler and M C Mozer. When does fading enhance perceptual category learning? *Journal of experimental psychology. Learning, memory, and cognition*, 39(4):1162–73, July 2013.