World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:4, 2007

Stochastic Learning Algorithms for Modeling Human Category Learning

Toshihiko Matsuka and James E. Corter

***Abstract*—**Most neural network (NN) models of human category learning use a gradient-based learning method, which assumes that locally-optimal changes are made to model parameters on each learning trial. This method tends to underpredict variability in individual-level cognitive processes. In addition many recent models of human category learning have been criticized for not being able to replicate rapid changes in categorization accuracy and attention processes observed in empirical studies. In this paper we introduce stochastic learning algorithms for NN models of human category learning and show that use of the algorithms can result in (a) rapid changes in accuracy and attention allocation, and (b) different learning trajectories and more realistic variability at the individual-level.

***Keywords*—** category learning, cognitive modeling, radial basis function, stochastic optimization.

I. Introduction

RECENT neural network (NN) models of classification learning, ALCOVE [1], RASHNL [2], and SUSTAIN [3] for example, share a number of common aspects, including multilayer architectures and learned dimensional attention weights as well as learned association weights between stimulus input nodes in the transformed feature space and the output category nodes. One of these common elements is the use of gradient-based learning algorithms to adjust both association weights and dimensional attention parameters. In this method, weights are adjusted based on discrepancies between a target training signal and the output activations on the current layer, by computing the gradient of the error function in the multidimensional parameter space. This is accomplished by taking the partial derivative of the error function with respect to each of the network parameters (weights) in turn. The algorithm then adjusts each of these weights proportionally to its partial derivative. This method is an effective means of finding optimal estimates for parameters, as long as the overall error function is not characterized by strong local minima. Thus the algorithm has normative orientation/justification, i.e., it models how people "should" learn or process information in terms of error minimization.

But, is the gradient method plausible *descriptively* (i.e., does it describe how people actually learn)? It seems implausible that people explicitly compute the gradient on each trial while attempting a categorization task. On the other hand, people's general learning mechanisms might have evolved so as to approximate gradient learning.

But in any case, we should first ask if the gradient-based learning algorithms are successful in replicating empirical data in human category learning. Results in the literature demonstrate that these methods have been successful in reproducing group learning curves [1 – 3]. However, recent studies in our lab suggest these models may underpredict variability in individual-level empirical data, particularly differences in attention allocation measures [4 – 6].

Another important issue raised by experimental psychologists in the field of cognitive modeling is whether learning happens gradually or in an all-nor-none fashion. Gradual learning curves have been reported for aggregated data and/or categorization tasks defined by complex concepts [1-3]. On the other hand, when individual subjects were considered, the learning curves for some participants change suddenly, following an all-or-none like learning pattern, particularly for very simple categorization tasks [7, 8]. Similarly, some empirical studies suggest that human's attention allocation to individual dimensions can change quite rapidly [8, 9]. Most cognitive models based on gradient-based learning mechanisms appear to have difficulty reproducing such rapid changes in attention and classification accuracy.

In the present research we (a) introduce alternative learning algorithms for NN models of classification learning, specifically stochastic learning algorithms based on simulated annealing, and then (b) test their descriptive validities in replicating empirical phenomena observed particularly at the individual level.

II. Stochastic Learning for Modeling Human Category Learning

A. Qualitative descriptions

Our proposed algorithm is based on a specific simulated annealing algorithm [10]. In the present algorithm, initial association weights are randomly selected from a uniform distribution centered at 0, and initial dimension attention weights are equally distributed across all dimensions. This equal attention allocation at the initial stage of learning is

Manuscript received December 9, 2003. This work was supported in part by the Nation Science Foundation under Grant EIA-0205178 and the James McDonnel Foundation.

T. Matsuka is with Rutgers University Mind and Brain Analysis (RUMBA) Laboratory, Rutgers University. 101 Warren Street, Newark, NJ 07102 USA (phone: +1 973-353-5440 x239; fax: +1 973-353-1170; e-mail: matsuka@psychology.rutgers.edu).

J. E. Corter is with Teachers College, Columbia University, New York, NY 10027 USA. (e-mail: jec34@columbia.edu)..

motivated by the results of empirical studies [4, 8], according to which, many subjects tended to evenly allocate attention to the feature dimensions in the early stages of learning. In other words, the model does not have any prior information or beliefs about which feature dimensions are more informative than the others as observed in real human subjects in laboratory experiments.

In our algorithm, at the beginning of each training epoch, a hypothetical "move" in the parameter space is computed by adjusting each parameter by an independently sampled term. These adjustment terms are drawn from a prespecified zero-mean symmetric distribution (e.g. Cauchy). The move (i.e., the set of new parameter values) is then accepted or rejected, based on the computed relative fit of the new parameter configuration. Specifically, if the new parameter values result in a better fit, they are accepted at the probability of one. If they result in a worse fit, they are accepted with some probability *P*. This probability is a function of a parameter called the "temperature", which decreases across blocks according to the annealing schedule. This particular annealing algorithm is relatively efficient, in that the adjustment in the network parameters is very rapid initially, and gradually decreases over learning blocks.

Our present model may be interpreted as a model that randomly generates a hypothesis and then evaluates it. In early stages of learning the present model is quite likely to produce "radical" hypotheses (i.e., the new set of hypotheses thus parameter values are very different from the currently valid hypotheses), and the probability of accepting a hypotheses set with worse utility could be relatively high. But, as learning progresses, the widths of the random distribution and *P* decrease, so that the model increasingly stabilizes its hypotheses and establishes more concrete concepts about the category. In other words, the model's concepts about the category evolve as learning progresses by permitting "good" hypotheses (and occasionally "bad" ones) to survive and using such enduring hypotheses as bases for generating a new set of hypotheses.

1) Key Characteristics

This learning algorithm can be applied to any feedforward neural network model of category learning. We assume that there is no (back) propagation of classification errors in the present model. Rather, we propose a very simple operation (i.e., comparison of two values) along with the operation of stochastic processes as the key mechanisms in human category learning.

1. Initial network association weights (***w***) are set to small random values, and initial dimension attention weights (***α***) are set equal across dimensions.

2. In learning, the attention strengths and association weights are updated with a random move in the parameter space, based on a prespecified zero-mean symmetric distribution (e.g., the Cauchy distribution).

3. If the new parameter configuration (or hypothesis) results in better categorization accuracy (based on a "mini-simulation" using the network model), then the hypothesis is accepted, and the new attention strengths and association weights replace the old values. In the case of a *decrease* in categorization accuracy due to the move, the hypothesis is accepted with some probability *P* ($0<P<1$).

4. *P* is relatively large in the early stages of learning, but it decreases as learning progresses. This decrease is associated with a decrease in a parameter called the "temperature", by analogy with the physical process that occurs as a metal cools.

Thus, the present model does not assume that learning is associated with monotonic increases in accuracy (and attention) or continuous search for better categorization processes by human. Rather, it models random fluctuations or "errors" in people's memory and learning processes, and how people utilize and "misutilize" such errors.

As a test of these ideas, we have embedded the present learning algorithm into the ALCOVE model [1].

III. ALCOVE

ALCOVE [1], for Attention Learning COVEring map, is an exemplar-based multi-layer adaptive network model of categorization based in part on the Generalized Context Model or GCM [11]. The first layer of ALCOVE is a stimulus input layer. Each dimension has an attention strength (α_i) associated with it. The next layer in the network is the exemplar layer. Each node in this layer corresponds to an exemplar, described by its position in the multidimensional stimulus space, and receives input from the input layer. The activation of each exemplar node is calculated based on its similarity to the presented stimulus:

$$h_j = \exp\left[-c\left(\sum_i \alpha_i \,|\, \psi_{ji} - x_i \,|\right)\right] \tag{1}$$

where ψ_{ji} is the value of exemplar node *j* on dimension *i*, x_i is the activation of input feature dimension *i*, *c* is a constant called the *specificity* that controls overall attention, and α_i is the attention strength for dimension *i*. In ALCOVE, the attention strengths essentially stretch and shrink dimensions.

The activity of the exemplar nodes is fed forward to the third layer, the category layer, whose nodes correspond to the categories being learned. The strength of association between category node *k* and exemplar node *j* is denoted by w_{kj}. The activation of category node *k* is then computed as the sum of weighted activations of all exemplars, or

$$y_k = \sum_j w_{kj} h_j \tag{2}$$

The probability that a particular stimulus is classified as category *k*, denoted as $P(k)$, is assumed equal to the activity of category *k* relative to the summed activations of all categories, where the activations are first transformed by the exponential function [1]:

$$P(K) = \frac{\exp(\phi y_k)}{\sum_k \exp(\phi y_k)} \tag{3}$$

where ϕ is a real-value mapping constant that controls decisiveness of classification responses.

The standard version of ALCOVE [1] uses a form of gradient descent for updating weights. The error term is defined as the sum of squared differences between the desired and the predicted outputs:

$$E = \tfrac{1}{2}\sum_k \varepsilon_k^2 = \tfrac{1}{2}\sum_k (t_k - y_k)^2 \tag{4}$$

Partial derivatives of the error function with respect to the association weights w_{kj} and the attention strengths α_i are used to compute the weight updates:

$$\Delta w_{kj} = -\lambda_w \frac{\partial E}{\partial w_{kj}} = \lambda_w \cdot \varepsilon_k \cdot h_j$$
$$\Delta \alpha_i = -\lambda_\alpha \frac{\partial E}{\partial \alpha_i} = -\lambda_\alpha \sum_j \left[\sum_k \varepsilon_k \cdot w_{kj}\right] h_j \cdot c\,|\psi_{ji} - x_i| \tag{6}$$

where λ_w and λ_α are the learning rates for the association weights and attention strengths, respectively. It is this gradient-based learning method that we propose to replace with the stochastic learning method.

IV. Stochastic Learning ALCOVE

Here, we have evaluated two applications of stochastic learning to ALCOVE: one version in which we implement stochastic learning for adjusting both dimensional attention weights and the network association weights (ALCOVE-CSL, for "completely stochastic learning"), and one in which stochastic learning is used to adjust only the dimension attention weights in ALCOVE (ALCOVE-SAL, for "stochastic attention learning "). Again, it should be noted that this learning algorithm is very general and can be applied to virtually any NN model of category learning.

A. ALCOVE-CSL algorithm

STEP 0: Initialize:

Problem specific parameters (T^0,υ)

T^0 : initial temperature.

υ : temperature decreasing rate

Association weights w_{kj},

$w_{kj} \sim U(MIN_w, MAX_w)$.

where MIN_w and MAX_w are minimum and maximum values for $\boldsymbol{w}$.

Attention strengths α_i,

$\alpha_i = 1/I^*(MAX_\alpha - MIN_\alpha) + MIN_\alpha$, for all $i = 1 \ldots I$, where I is the number of feature dimensions.

Exemplar ψ_{ji}

$\psi_{ji} = x_{*i}$,

where subscript * indicates unique patterns.

STEP 1: Calculate output activations

$$O_k = \sum_J w_{kj} \exp\left[-c\left(\sum_I \alpha_i |\psi_{ji} - x_i|\right)\right] \tag{7}$$

STEP 2: Calculate fit index for one training block:

$$F(\theta^t) = \sum_{n=1}^{N}\sum_{k=1}^{K} (d_k - O_k)^2 \tag{8}$$

where w_{kj}, $\alpha_i \in \theta$, K = # categories, N = # input in one block, d_k is a desired output for category node k.

STEP 3: Accept all weight and attention parameters ($\boldsymbol{\alpha}$ & $\boldsymbol{w}$) at the probability of:

$$P(\theta^t \mid \theta^A) = \left\{1 + \exp\left(\frac{F(\theta^t) - F(\theta^A)}{T^t}\right)\right\}^{-1} \tag{9}$$

if $F(\theta^t) > F(\theta^A)$, or 1 otherwise, where $F(\theta^A)$ is the fit index for the previously accepted parameter set.

STEP 4: Reduce temperature:

$$T^t = T^o \delta(-\upsilon, t) \tag{10}$$

where δ is the temperature decreasing function that take temperature decreasing rate, υ, and time t as inputs.

STEP 5: Generate new $\boldsymbol{w}$ and $\boldsymbol{\alpha}$

$$w_{kj}^t = w_{kj}^A + y(MAX(w) - MIN(w))$$
$$\alpha_i^t = \alpha_i^A + y(MAX(\alpha) - MIN(\alpha)) \tag{12}$$

where

$$y = \mathrm{sgn}(u - 0.5) T\left[\left(1 + 1/T\right)^{|2u-1|} - 1\right] \tag{13}$$

Here, u indicates a random number drawn from the Uniform distribution.

REPEAT Steps 1~5 until criterion is met

B. ALCOVE with SAL

Stochastic Attention Learning (SAL) incorporates both gradient and stochastic method for learning. In particular, SAL updates its association weights using a gradient method (Equation 5), and attention strengths by the stochastic learning method, i.e., (12) & (13).

Since SAL incorporates gradient learning for its association weights, the badness-of-fit index at time t is often less than that

at time *t-1*, even with an "inappropriate" random movement in attention allocation. In other words, the present algorithm as described in the previous section would accept many useless moves for attention distribution, particularly in the early stages of learning. However, this seems both unnecessary and inefficient. Thus, we modified SAL to include a threshold parameter ζ, which controls for the probability of accepting new attention weight values, to make the model accept only moves that satisfy a prespecified criterion (i.e., above the threshold) for categorization accuracy.

Thus for SAL, (9) becomes

$$P(\alpha^t \mid \alpha^A) = \left\{1 + \exp\left(\frac{F(\alpha^t) + \zeta - F(\alpha^A)}{T^t}\right)\right\}^{-1} \quad (14)$$

if $\{F(\alpha^t)+\zeta\} > F(\alpha^A)$, or 1 otherwise, where

$$F(\alpha^t) = \sum_{n=1}^{N}\sum_{k=1}^{K}(d_k^t - O_k^t)^2 \quad (15)$$

C. Comparisons to the previous learning algorithm

1) Gradient decent vs. Stochastic Learning

There are several apparent differences between gradient-based and stochastic learning models. Two differences relevant to cognitive modeling are (a) the rate of learning, and (b) utilization of collinear diagnostic dimensions. While learning curves for a cognitive model with a gradient decent method gradually change, a model with the stochastic learning methods could learn correct category membership either gradually, suddenly (e.g. all-or-none fashion), or in a combined manner, depending on the configurations of its free parameters (e.g., temperature decreasing rate). Another related issue is that while a gradient decent method can be considered as a learning process searching for the new parameter set in the most effective direction at any given time (and with given information on hand), this stochastic learning may be considered as learning process characterized by (pseudo) random exploration of the stimulus space.

When some diagnostic feature dimensions are highly collinear or correlated (i.e., flat minima), a gradient method would pay good amounts of attention to those collinear diagnostic dimensions, and the amounts of attention allocated to those dimensions would be similar. Whereas, for the stochastic learning methods, many patterns of attention distribution can be expected for such loosely defined input-output relationship, in which many parameter configurations (i.e., association weights and attention distribution) can result in acceptable levels of classification accuracy. Although it is not well understood how real humans would utilize or allocate attention to diagnostic collinear dimensions, one recent study [6], as discussed in next section, reported that people tend to allocate rather different amounts of attention to the collinear diagnostic dimensions, showing some individual differences.

2) RULEX vs. Stochastic Learning

One unique model of categorization utilizing learning algorithms other than gradient methods is RULEX [12]. In particular, it incorporates a sequential hypothesis-testing-like learning algorithm. In its first stage of learning, RULEX tries to identify a categorization rule defined by a single perfectly predictive feature dimension, and the search process continues until all feature dimensions are tested. In the second stage, it searches for (multiple) imperfect single dimension rules, followed by conjunctive rules in the third stage.

Although, RULEX sounds plausible and it could replicate individual differences for several stimulus sets, it may be considered to be more normative than real humans in term of efficiency in information usage. That is, RULEX predicts that people are *always* capable of identifying the minimal and sufficient numbers of diagnostic feature dimensions, and use only exactly the same diagnostic dimensions once they learned.

Our present learning models' take-all-or–none parameter updating strategy may be considered as a type of hypothesis testing learning model, which makes it similar to RULEX. However, its random search method, interpreted as unstructured hypothesis generation and search, is very distinct from RULEX whose hypothesis generation algorithm is very strategic and well structured. Our present models can pay attention to any numbers of feature dimensions as long as the parameter configurations result in the acceptable categorization accuracies. Thus, the stochastic learning methods can result in paying attention to irrelevant dimensions as long as its parameter configuration is associated with good classification accuracy. This in turn may lead to "superficial" beliefs that some irrelevant dimensions are somewhat relevant to particular category concepts, apparently a common phenomenon in ordinary life (e.g., believing in jinx). In other words, when there are several minima, which is probably true for real world category learning task, stochastic learning can result in several different learning trajectories and parameters configurations (i.e., association weight & attention allocation), corresponding to possible individual differences. In contrast, RULEX would predict that people always pay attention to the least number of dimensions throughout the entire learning phase (i.e., if there is only one diagnostic dimension, it would never pay attention to more than one dimension in its entire learning process, because it starts making hypothesis that only a single dimension is diagnostic), which may be a too normative prediction in terms of efficiency of information usage.

V. Simulations

In order to evaluate the ability of the stochastic learning algorithms to account for human data on classification learning, we conducted three simulation studies. In Simulation 1, we tested if the stochastic learning can replicate rapid changes in classification accuracy and attention allocation in category learning for a single simulated subject. In Simulation 2, we simulated the results of a recent empirical study on classification learning [5, 6] to see if the algorithm can reproduce individual differences in attention processes in a

classification task involving with stimuli characterized by highly collinear feature dimensions. In Simulation 3, we examined if the algorithms accurately reproduce aggregated learning curves.

The simulations reported below compare several ALCOVE-based models. The main comparison we are interested in is to compare the performance of standard ALCOVE with ALCOVE incorporating stochastic attention learning (ALCOVE-SAL), and ALCOVE incorporating completely stochastic learning (ALCOVE-CSL). However, for Simulation 2, we also investigate if individual differences could be otherwise accounted for within standard gradient-based ALCOVE. To do this, we also tried another way (besides stochastic learning) of handling random individual differences within the ALCOVE model, namely by randomly varying individual learning rates. This version of standard gradient-learning ALCOVE is referred to here as ALCOVE-RLR (Randomize Learning Rate).

A. Simulation 1: Rapid changes in accuracy and attention

In the present simulation study, we examined if stochastic learning algorithms can replicate rapid changes in classification accuracy and attention allocation as observed in some empirical studies [7 - 9]. Here, we used the simplest stimulus structure (T1) of Shepard, Hovland and Jenkins' stimulus sets [13]. Table I shows schematic representation of the stimuli used in the present simulation (i.e., T1). For T1 stimulus set, only Dimension 1 (D1) is necessary and sufficient for perfect classification. The main goal of Simulation 1 is to investigate which learning algorithm is capable of reproducing psychological phenomena observed in Rehder and Hoffman [8], as described below, that used the same T1 stimulus set.

There are two important observations reported by Rehder and Hoffman [8] that can help assess descriptive validity of cognitive models. First, the study showed that people initially pay attention to all three feature dimensions, and then learn to allocate attention almost exclusively to the diagnostic dimension. As Rehder and Hoffman [8] claimed, the results casts doubt about the descriptive validity of RULEX, which predicts people would *initially* pay attention to a single dimension or have and test a hypothesis that there was only one diagnostic dimension. Second, when individual data were analyzed, sudden changes were observed for classification accuracies and attention distribution. More specifically, there were multiple plateaus in learning curves, suggesting that participants follow an all-or-none type of learning for this particular task. This in turn casts doubt on backpropagation's (i.e., gradient decent) gradual learning as a descriptive model of human category learning. It should be noted, however, that this categorization task was very easy, and all-or-none type of learning trajectory may not be observed with more complex categorization tasks.

1) Simulation Method

Three ALCOVE-type models of category learning were evaluated in the present simulation studies, namely the standard ALCOVE, ALCOVE-SAL, and ALCOVE-CSL. They were run in a simulated training procedure to learn the correct classification responses. ALCOVE was run for 50 blocks of training, where each block consisted of a complete set of the training instances, while ALCOVE-SAL and ALCOVE-CSL were run for 150 blocks. For each model, the gradient or rate of change in attention allocated to Dimension 1 was calculated by subtracting the amount of attention allocated to Dimension1 at time *t*-1 from that of time *t*. This measure was used as an index of how rapidly attention distributions changed.

TABLE I. SCHEMATIC REPRESENTATION OF STIMULUS SET USED IN SIMULATIONS 1 AND 3.

Stimulus			Category					
D1	D2	D3	T1	T2	T3	T4	T5	T6
1	1	1	A	B	A	A	A	A
1	1	0	A	B	A	A	A	B
1	0	1	A	A	A	A	A	B
1	0	0	A	A	B	B	B	A
0	1	1	B	A	B	A	B	B
0	1	0	B	A	A	B	B	A
0	0	1	B	B	B	B	B	A
0	0	0	B	B	B	B	A	B

2) Results

The results of one simulated subject for each model are shown in Fig. 1: The models' predicted classification accuracies, relative attention allocations to the three dimensions, and rates of change in attention allocated to Dimension 1, are plotted in the top, middle bottom row, respectively. Note that classification accuracies and attention distributions for accepted hypotheses (i.e., parameter configuration) were plotted.

All three models learned to allocate the highest amount of attention to Dimension 1 and learned to ignore or pay less attention to Dimensions 2 and 3. The rate of attention change for ALCOVE was very smooth and its magnitude was much smaller than those of ALCOVE-SAL and ALCOVE-CSL. ALCOVE-SAL and ALCOVE-CSL produced oscillating graphs (Fig. 1E &1I) with higher magnitudes of change, showing sudden shifts in attention distributions. In addition, both ALCOVE-SAL and CSL also showed sudden changes in classification accuracies (Fig. 1D & 1G), replicating learning patterns observed in Rehder and Hoffman [8]. Note that for both stochastic learning ALCOVEs, the sudden changes in attention distributions and classifications accuracies were well synchronized. Although ALCOVE-SAL and CSL replicated rather large sudden changes in attention allocation and classification accuracies, the learning curves changed somewhat gradually after the sudden changes.

In sum, these results support the descriptive validity of our proposed stochastic learning algorithms, showing capabilities of reproducing rapid changes in attention allocation and classification accuracy observed in some empirical studies.

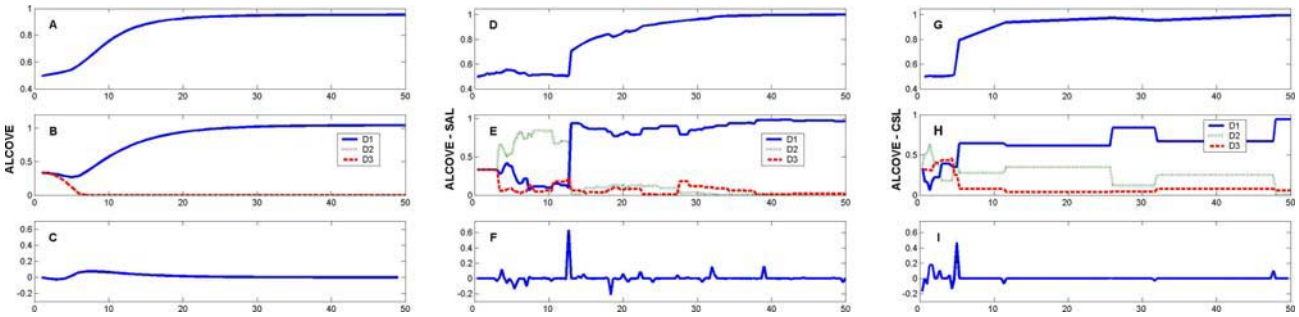

Figure 1. The results of Simulation 1. A: predicted classification accuracy by ALCOVE, B: predicted relative attention allocation to the three feature dimension by ALCOVE; C: predicted rate of change in attention allocated to Dimension 1 by ALCOVE; D - E: predictions by ALCOVE-SAL; G - I by ALCOVE-CSL.

B. Simulation 2: Individual Differences

In this simulation study, we examined how the models account for individual differences in attention learning for categories defined by highly collinear feature dimensions. To do this, we simulated the results of an empirical study on classification learning, Study 2 of Matsuka [5]. In this study, there were two perfectly redundant feature dimensions, Dimensions 1 & 2 (see Table II), and those two dimensions were also perfectly correlated with the category membership. Thus, information from only one of the two correlated dimensions was necessary and sufficient for perfect categorization performance. Besides classification accuracy, data on the amount of attention allocated to each feature dimension was collected in the empirical study. The measures of attention used were based on feature viewing time, as measured in a MouseLab-type interface [14].

To summarize the empirical results that we are trying to simulate, 13 out of 14 subjects were able to categorize the stimuli almost perfectly (Fig. 2, Top right panel), and *on average* subjects paid attention to both of the correlated dimensions approximately equally (Fig. 2, Top middle panel). When Matsuka and Corter [5] analyzed attention data at an individual-level, they found that many subjects tended to pay attention primarily to only one of the two correlated dimensions, particularly in the late learning blocks (Fig. 2, Top right panel). This suggests that they tend to utilize the minimal necessary information for this task.

TABLE II: STIMULUS STRUCTURE USED IN STUDY 2 OF MATSUKA (2002)

Category	Dim1	Dim2	Dim3	Dim4
A	1*	1*	3	4
A	1*	1*	4	1
A	1*	1*	1	2
B	2*	2*	2	1
B	2*	2*	3	2
B	2*	2*	4	3
C	3*	3*	1	3
C	3*	3*	2	4
C	3*	3*	3	1
D	4*	4*	4	2
D	4*	4*	2	3
D	4*	4*	1	4

*Diagnostic feature

1) Simulation method

There were four ALCOVE-type models involved in the present simulation study, namely standard ALCOVE, ACLOVE with random learning rate (ALCOVE-RLR), ALCOVE-SAL, and ALCOVE-CSL. The final parameter values used for ALCOVE were chosen by a simulated annealing method [6, 10] to minimize the objective function (i.e., sum of squared error) in reproducing the classification accuracies by human subjects. The same free parameter values for ALCOVE were used for ALCOVE-RLR, except for its learning rate for attention. For each simulated subject, its attention learning rate was selected from the uniform random distribution with MIN = 0.00255, MAX = 0.0102, which were a half and twice the value of attention learning rate selected by the optimization process for the standard ALCOVE in the present simulation study, respectively. For ALCOVE-SAL and CSL, the final parameter values were selected arbitrary based on the values identified for the standard ALCOVE.

The four models were run in a simulated training procedure to learn the correct classification responses for the stimuli of the experiment. ALCOVE and ALCOVE-RLR were run for 48 blocks of training, where each block consisted of a complete set of the training instances, while ALCOVE-SAL and ALCOVE-CSL were run for 500 blocks. For each model, the final results are based on 50 replications.

2) Results

Fig. 2 summarizes the findings from the simulation study. The top row of this figure shows the empirical data from Study 2 of Matsuka [4] (also reported in [5]), including the learning curve for overall classification accuracy (left panel), the attention learning curves (middle panel), and relative attention allocated to Dimensions 1 vs. 2 (the redundant diagnostic dimensions) in the late training blocks (i.e., the last half).

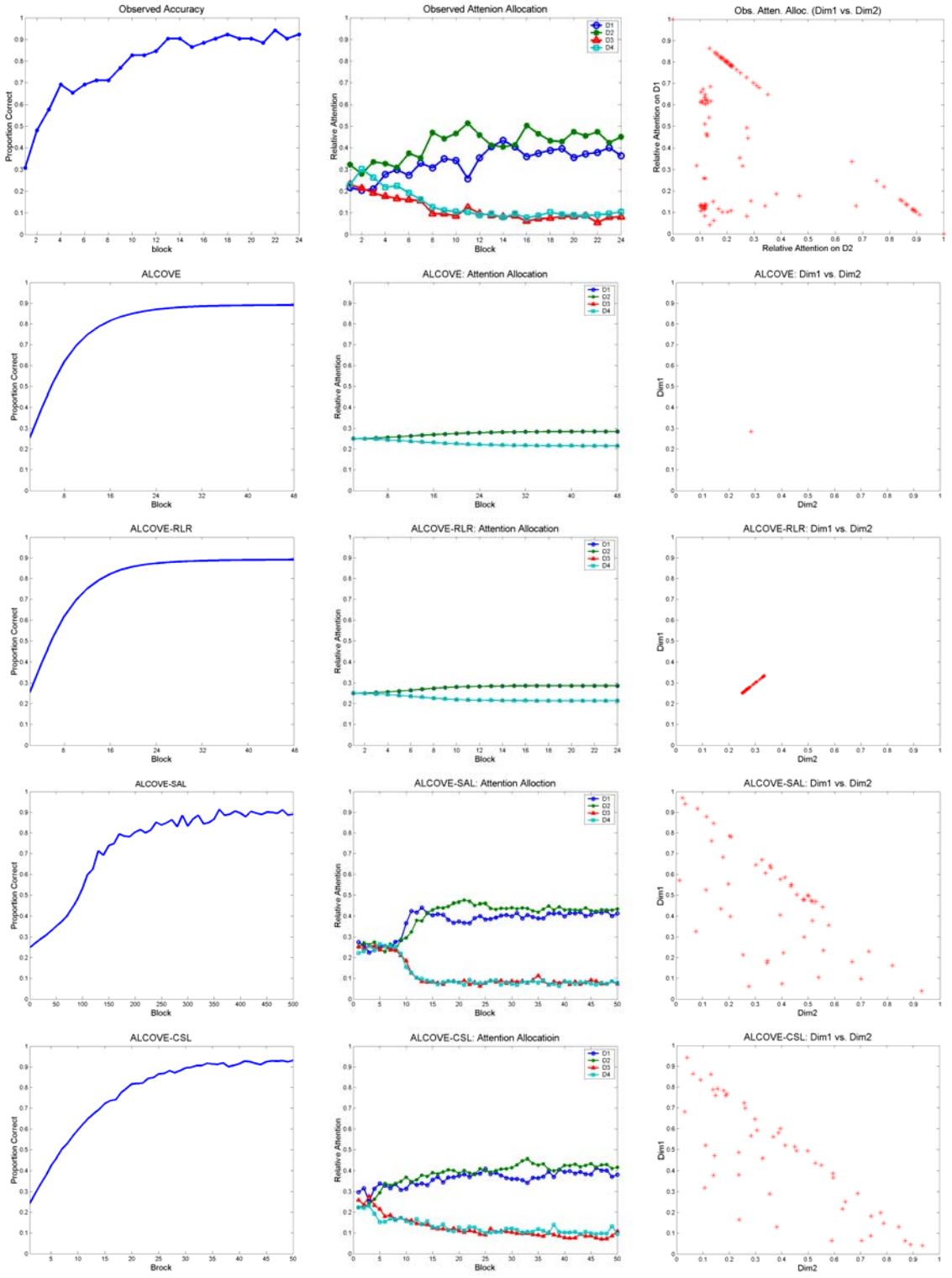

Figure 2: Results of Simulation 2. The first row shows observed learning curve (left panel), observed dimensional attention allocation (middle panel), observed attention allocation to Dimension 1 (Y-axis) and 2 (X-axis) in the last half of the training block (right panel). The predictions by ALCOVE, ALCOVE-RLR, ALCOVE-SAL, and ALCOVE-CSL are shown in the 2nd, 3rd, 4th and 5th row, respectively.

Results for each of the models are shown in the remaining rows. Standard ALCOVE learned to allocate attention equally to the two diagnostic dimensions, but showed almost no inter-individual variability in classification accuracy nor attention. In addition, it produced no intra-individual variability in the amount of attention allocated to Dimension 1 and 2 (Fig. 2, Second row, last column). In other words, in ALCOVE, every simulated subject paid exactly the same

amount of attention to the two diagnostic redundant dimensions.

ALCOVE-RLR, using random learning rate for attention strengths, but still using the standard gradient learning algorithm, showed some inter-individual variability. However, this model, too, predicted that every simulated subject would pay exactly the same amount of attention to the redundant dimensions. That is, the relative amounts of attention allocated to Dimension 1and 2 were directly proportional (Fiuge2, third row, last column).

ALCOVE-SAL, the version of ALCOVE modified to incorporate stochastic learning of attention weights, showed much more variability among subjects in attention allocation, more closely resembling the empirical data. ALCOVE-CSL gave similar results, but exhibited some minor differences from ALCOVE-SAL.

In sum, the stochastic learning models were shown to be capable of reproducing individual differences/preferences on attention distributions to two feature dimensions that contain exactly the same information. On the other hand, the gradient-based learning always used information from the two dimensions equally, not being able to reproduce the key psychological phenomenon in the present study.

C. Simulation 3. Replication of Nosofsky et al. (1992)

Thus far, we have shown that our proposed stochastic learning algorithms are successful for reproducing individual-level data (i.e. rapid change & individual differences in attention processes). However, we have not explicitly tested the algorithms' capabilities of reproducing aggregated data. In the present simulation study, we simulated a classical study of categorization [13] which is often used as a benchmarking stimulus set [15]. The stimulus structures are shown in Table I. The results of previous empirical studies showed that Type 1 (T1) was the easiest to learn to classify, followed by T2, T3, T4, T5, and T6 being the most difficult. More precisely, Nosofsky et al. [15] showed that the numbers of error made (i.e., classification difficulties) for those stimulus structures were significant except T3, T4, and T5.

1) Simulation method

In the present study, we tested only ALCOVE-SAL and CSL, as the standard ALCOVE has previously been shown to be able to replicate the Shepard et al. results [15]. The two models were run in a simulated training procedure to learn the correct classification responses for the stimuli. ALCOVE-CSL was run for 250 blocks of training, where each block consisted of a complete set of the training instances, while ALCOVE-SAL was run for 150 blocks. For each model, the final results are based on 500 replications.

2) Results

Fig. 3 summarizes the findings from the Simulation 3. Both ALCOVE-CSL and SAL were able to reproduce the order of difficulty successfully. That is ALCOVE-CSL and SAL find T1 to be the easiest, followed by T2, T3, T4, T5, and T6.

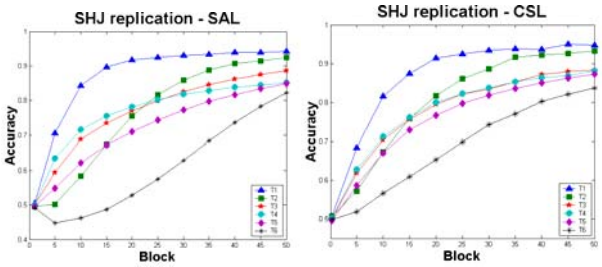

Figure 3. Results of Simulation 3. Both ALCOVE-SAL and ALCOVE-CSL were able to reproduce the order of difficulty successfully reported in Shepard et al. [13] & Nosofsky et al. [15].

VI. Discussion and Conclusion

Here we have investigated the possibility of using stochastic learning rather than gradient-based methods in neural network models of human classification learning. In the present simulations we have explored the effectiveness of this method in several variants of the ALCOVE-type model [1]. Our main goals were to see if stochastic learning algorithms 1) were able to replicate rapid change in classification accuracy and attention processes, and 2) offered a better account of individual differences in classification accuracy learning curves and in final distribution of attention, particularly distribution of attention to two perfectly correlated dimensions. The simulation studies showed that the new algorithms are satisfactory in these regards.

Stochastic learning algorithms have other desirable properties as well. It could be argued that stochastic learning may be more psychologically plausible than gradient-based methods, which require more mental effort and assume that optimal adjustments are made to the vector of parameters on each trial. One caveat to these results is that the stochastic learning algorithms learn more slowly than the standard gradient methods in categorization tasks with relatively small (both number of stimulus feature dimensions and number of unique exemplars) and well-defined stimulus sets that are usually used in laboratory experiments. However, for more realistic category learning involving complex category structures and/or stimuli with many feature dimensions, stochastic learning may be able to learn faster than ordinary gradient type learning.

A. Application for GECLE modeling

Matsuka [16] introduced a framework for modeling human category learning named GECLE, based on radial basis functions. One main strength of their modeling approach is the flexibility of its attention mechanisms, namely the use of the Mahalanobis distance function (which is able to pay attention to correlated dimensions) for calculating similarity to exemplars, localized receptive fields (i.e., each exemplar or prototype can have a uniquely shaped and oriented receptive field), and a variety of choices in the activation transfer

function. However, since the model uses a gradient method for updating association weights, attention strengths, and locations of reference points (i.e., exemplars or prototypes), its activation transfer function is constrained to be a differentiable function. However, the use of our derivative-free stochastic learning algorithm can eliminate the constraints and make the model more flexible.

B. Distribution of random numbers

In the present research, the random moves in parameter space were drawn from the Cauchy distribution, mainly because its fatter tails are more likely than the Gaussian distribution to produce the rapid and/or large shift in attention allocation that has been reported by some empirical studies. However, with a proper experimenter-defined parameter setting (e.g., initial temperature & temperature decreasing rate), such shifts in attention might have been achieved with the Gaussian distribution. Moreover, it may be possible that shifts could be drawn from other types of distributions, including rectangular, skewed, or multi-modal distributions. Further simulation and empirical studies seem useful for investigating this issue.

ACKNOWLEDGMENT

Authors thank Stephen Jose Hanson, Catherine Hanson, Yasuaki Sakamoto, Areti Chouchourelou, and researchers at RUMBA for their helpful comments and suggestions.

REFERENCES

[1] Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning, *Psychological Review, 99.* 22-44.
[2] Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083-1119.
[3] Love, B. C. & Medin, D. L. (1998). SUSTAIN: A model of human category learning. *Proceeding of the Fifteenth National Conference on AI (AAAI-98),* 671-676.
[4] Matsuka, T. (2002). Attention processes in computational models of category learning. Unpublished doctoral dissertation. Columbia University, New York, NY.
[5] Matsuka, T. & Corter, J. E. (2003). Empirical studies on attention processes in category learning. Poster presented at 44th Annual Meeting of the Psychonomic Society. Vancouver, BC, Canada.
[6] Matsuka, T., Corter, J. E. & Markman, A. B. (2003). Allocation of attention in neural network models of categorization. Under review
[7] Bower, G. H. & Trabasso, T. R. (1963). Reversals prior to solution in concept identification. *Journal of Experimental Psychology, 66*, 409-418.
[8] Rehder, B. & Hoffman, A. B. (2003). Eyetracking and selective attention in category learning [CD-ROM]. Proceedings of the 25th Annual Meeting of the Cognitive Science Society, Boston, 2003.
[9] Macho, S. (1997). Effect of relevance shifts in category acquisition: A test of neural networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 30-53.
[10] Ingber, L. (1998). Very fast simulated annealing. *Journal of Mathematical Modelling, 12:* 967-973.
[11] Nosofsky, R. M. (1986). Attention, similarity and the identification –categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57
[12] Nosofsky, R. M., Palmeri, T. J., McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79
[13] Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classification. *Psychological Monograph, 75* (13).
[14] Bettman, J. R., Johnson, E. J., Luce, M. F., Payne, J. W. (1993). Correlation, conflict, and Choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 931-951.
[15] Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition, 22,* 352-369.
[16] Matsuka, T. (In press). Generalized exploratory model of human category learning. *International Journal of Computational Intelligence*.