Biased stochastic learning in computational model of category learning

Toshihiko Matsuka (matsuka@psychology.rutgers.edu)
RUMBA, Rutgers University – Newark
101 Warren St., Smith Hall 327, Newark, NJ 07102 USA

Abstract

Matsuka and Corter (2003b) presented evidence that people tend to utilize only the minimally necessary information for classification tasks. This approach for categorization was efficient and valid for the stimulus set used in the experiment, but might be considered a statistically or mathematically non-normative approach. In the present paper, I hypothesized human category learning processes are biased toward simpler representation and/or conception rather than complex but normative ones. In particular, a few variants of "biased" learning algorithms are introduced and applied to Matsuka and Corter's stochastic learning algorithm (2003a, 2004). The result of a simulation study showed that the biased learning models account for empirical results successfully.

Introduction

In their recent work, Matsuka and Corter (2003b & 2004) investigated the possibility of using stochastic learning rather than gradient-based methods in neural network models of human category learning. They introduced stochastic learning models to more accurately account for human category learning. The gradient based learning algorithm used in many neural network models may be considered to have a normative justification (i.e., it models how people "should" learn or process information), but may not be descriptively valid at the individual level. Models utilizing a gradient method for learning seem to require a high degree of mental effort and assume that optimal adjustments are made to the vector of parameters on each trial. In contrast, Matsuka & Corter's stochastic learning model (2003a, 2004) does not assume that learning is associated with monotonic increases in accuracy (and attention) or continuous search for better categorization processes by humans. Rather, it models random fluctuations or "errors" in people's memory and learning processes, and how people utilize and "misutilize" such errors.

In their simulation studies (Matsuka & Corter 2004a), the effectiveness of stochastic learning methods applied to an ALCOVE-like model (Kruschke, 1992) was evaluated in several settings. The modified models were shown to be satisfactory in replicating two phenomena observed in empirical studies on categorization; namely, rapid change in attention processes (Macho 1997; Rehder and Hoffman 2003), and individual differences in distribution of attention (Matsuka & Corter 2003b).

Although the stochastic learning model reproduced more realistic individual differences than models with a gradient type learning algorithm, it did not replicate one tendency observed in the empirical study of Matsuka and Corter (2003b). They found that for four dimensional stimulus sets with two diagnostic but perfectly correlated dimensions, the proportion of human participants who paid attention primarily to only one of the two correlated dimensions was higher than that of those who paid attention to both of the two correlated dimensions approximately equally (see Figure 2, top row, third column). In other words, many participants utilized only the minimal necessary information for this task. In contrast, the stochastic learning model inadequately predicted that a higher proportion of participants would pay attention to the two correlated dimensions approximately equally.

The strategy of using minimal information may be a very natural and efficient usage of limited mental resources for humans. This would be particularly true for real world categorization tasks, where the number of feature dimensions could easily exceed a manageable number, in which many are not necessary or crucial (e.g., irrelevant and or highly correlated) for successful categorization. There are several ways that could lead people to use a lesser amount of information, resulting in simple conception of categories. One possible explanation is that there may be an implicit or explicit penalizing mechanism in human cognition that encourages less complete but simpler concepts than more complete but more complex concepts. Another possible explanation is that there may be a mechanism in human cognition that leads to a more thorough search for simple concepts.

In the present research, based on these remarks, I hypothesize and model human category learning as being biased toward simpler and heuristic concepts [1] (or representation) than complex and complete ones.

Biased Stochastic Learning

The proposed algorithm is based on a simulated annealing algorithm (Kirkpatrick, Gelatt, & Vecchi, 1983; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1958) and somewhat resembles Boltzmann Machine (Hinton & Sejnowski, 1986). 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 in the early stages of learning is motivated by the results of empirical studies (Matsuka, 2002; Rehder & Hoffman, 2003) that showed many participants initially tended to evenly allocate attention to the feature dimensions. In the present algorithm, at the beginning of each training epoch, a hypothetical "move" in

[1] In the present paper, the concepts of categories correspond to the configurations of the association weights and dimensional attention attractiveness.

the parameter space is computed by adjusting each parameter by an independently sampled term. These adjustment terms are drawn from a prespecified distribution. The move (i.e., the set of new parameter values) is then accepted or rejected, based on the computed relative fit or utility (defined below) of the new values. Specifically, if the new parameter values result in a better fit/utility, they are accepted. If they result in a poorer fit/utility, 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.

Because of the human's biased cognitive processes, possibly as a consequence of our implicit or explicit biased processes and/or preference toward simpler but less complete concept (these processes are discussed in detail in the model section), the learned concepts of categories, thus the configuration of the association weights and attention strengths, are inclined toward simpler ones. Note that in the present algorithm the notion of simplicity (or complexity) is directly related to the number of effective (non-zero, or non-subzero) association weights and attention strengths.

The proposed models would not require computation intensive (back) propagations of classification errors. Rather, in the present biased stochastic learning model framework, a very simple operation (e.g., comparison of two values) along with the operation of stochastic processes are assumed to be the key mechanisms in category learning. These learning algorithms can be applied to virtually any feed-forward NN model of human category learning

General Algorithm for Stochastic Learning

A general framework for the stochastic learning algorithm is discussed in this section. Here, the stochastic learning algorithm is embedded into ALCOVE, which is one of the most studied and applied computational models of category learning incorporating a selective attention mechanism (Kruschke, 1992). Again, it should be noted that this learning algorithm is very general and can be applied to virtually any NN model of category learning.

STEP 0: Initialization:

Problem specific parameters: (T^0,υ)

T^0 : initial temperature.

υ : temperature decreasing rate

Association weights w_{kj},, Attention strengths $\alpha_{i,}$, Exemplar ψ_{ji}

STEP 1: Calculate ALCOVE output activations:

$$O_k = \sum_J w_{kj} \exp\left[-c(\sum_I \alpha_i \mid \psi_{ji} - x_i \mid) \right] \qquad \text{(SL-1)}$$

STEP 2: Calculate fit index for the current parameter set:

$$F(w^t, \alpha^t) = \sum_{n=1}^{N} \sum_{k=1}^{K} (d_k - O_k^t)^2 \qquad \text{(SL-2)}$$

where *K* = # categories, *N* = # input in one block, d_k is a desired output for category node *k*. Here, the superscript *t* indicates time.

STEP 3: Accept or reject of parameter set, $\boldsymbol{\alpha}$ & $\boldsymbol{w}$:

Accept all weight and attention parameters at the probability of:

$$P(w^t, \alpha^t \mid w^s, \alpha^s, T^t) = \left\{ 1 + \exp\left(\frac{F(w^t, \alpha^t) - F(w^s, \alpha^s)}{T^t} \right) \right\}^{-1} \qquad \text{(SL-3)}$$

if $F(w^t, \alpha^t) > F(w^s, \alpha^s)$, or 1 otherwise, where $F(w^s, \alpha^s)$ is the fit index for the previously accepted parameter set, and T^t is temperature at time *t*.

STEP 4: Reduce temperature:

$$T^t = T^o \delta(\upsilon, t) \qquad \text{(SL-4)}$$

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

STEP 5: Generate new w_{kj} and α_i.

$$w_{kj}^t = w_{kj}^s + r^w, \; r^w \sim \Phi^w(\cdot) \qquad \text{(SL-5)}$$

$$\alpha_i^t = \alpha_i^s + r^\alpha, \; r^\alpha \sim \Phi^\alpha(\cdot) \qquad \text{(SL-6)}$$

where r^w and r^α are random numbers generated from prespecified distributions Φ^w and Φ^α.

REPEAT STEPS 1~5 until stopping criterion is met.

Biased Stochastic Learning Models

There are several approaches to model biased learning processes using stochastic learning model. Here, two simple approaches are introduced. The first biased learning model is based on the parameter regularization in which complex parameter configurations are penalized. The second model based on asymmetric random distributions, searches simpler parameter configurations more thoroughly.

Model 1: Bias via penalizing fitness function

In the present algorithm the utility index rather than the fit index is used for the decision on acceptance and rejection of the current parameter set. The utility of a particular parameter configuration is defined as a weighted sum of the accuracy in classification and the mental effort required by the parameter configuration. Thus, the utility index consists of two independent indices, namely "classification accuracy", *L* and "mental effort", *Q*, both dependent on learnable parameters $\boldsymbol{w}$ and $\boldsymbol{\alpha}$ at time *t*.

$$U(w^t, \alpha^t) = L(w^t, \alpha^t) + Q(w^t, \alpha^t) \qquad \text{(M1-1)}$$

The *L* function can be the same function for the fitness index (i.e., Eq. SL-2). Here, the *Q* function may be considered as a penalty function, penalizing "complex" parameter configurations that are believed to require more mental effort. The general form of *Q* function is given as follows:

$$Q(w^t, \alpha^t) = \gamma_w \phi^w(w_m^t) + \gamma_\alpha \phi^\alpha(a_m^t) \qquad \text{(M1-2)}$$

where ϕ^w and ϕ^α are functions calculating mental effort required for specific parameter configurations at time *t* (i.e., $\boldsymbol{w}^{\mathbf{t}}$ and $\boldsymbol{\alpha}^t$), and γ_w and γ_α are coefficients weighting these mental efforts. Note that γ_w and γ_α also control relative importance of *L* and *Q* functions (i.e., accuracy vs. simplicity). That is the hypothetical coefficient, γ_Q, weighting importance of *Q* function relative to *L* function is included in γ$_w$ and γ$_\alpha$. I.e., $\gamma_w = \gamma_Q \gamma_w^*$ and $\gamma_\alpha = \gamma_Q \gamma_\alpha^*$. Thus Equations M1-1 and M1-2 may be rewritten as:

$$U\left(w^t, \alpha^t\right) = L\left(w^t, \alpha^t\right) + \gamma_Q Q^*\left(w^t, \alpha^t\right) \qquad \text{(M1-1R)}$$

$$\gamma_Q Q^*\left(w^t, \alpha^t\right) = \gamma_Q \gamma_w^* \phi^w\left(w_m^t\right) + \gamma_Q \gamma_\alpha^* \phi^\alpha\left(a_m^t\right) \qquad \text{(M1-2R)}$$

There are several functions applicable for ϕ:

$$\phi^w = \sum_j \sum_k w_{kj}^2 \quad \text{(M1-3a)}; \qquad \phi^\alpha = \sum_i \alpha_i^2 \qquad \text{(M1-3b)}$$

$$\phi^w = \sum_j \sum_k I\left(\left|w_{kj}\right| > \zeta^w\right) \qquad \text{(M1-4a)}$$

$$\phi^\alpha = \sum_i I\left(\alpha_i > \zeta^\alpha\right) \qquad \text{(M1-4b)}$$

where ζ^w and ζ^α are threshold values, and *I*(*expression*) is the indicator function that returns 1 if the *expression* is satisfied. Equations M1-3a and M1-3b, often referred to as ridge penalty function or weight decay, encourage parameter settings that have small parameter values, whereas Equations M1-4a and M1-4b encourage parameter settings that have large number of parameters with less than the threshold values ζs. More general ϕ function is given as follows:

$$\phi^w = \sum_j \sum_k \frac{\left(w_{kj}/q\right)^2}{1 + \left(w_{kj}/q\right)^2}, \qquad \phi^\alpha = \sum_i \frac{\left(\alpha_i/q\right)^2}{1 + \left(\alpha_i/q\right)^2} \qquad \text{(M1-5)}$$

where q, which can be either time dependent or independent, controls types of penalization or encouragement. That is, Equations M1-5 approach Equations M1-3s as $q \to \infty$, and approach Equations M1-4s as $q \to 0$ (Cherkassky & Mulier, 1997).

In many simulation studies, relative, but not absolute predicted attention allocation strengths are analyzed and compared (e.g. Matsuka, 2002). In such cases, the relative attention strengths $a_i = \alpha_i/\Sigma(\alpha_m)$ should be used as inputs for the penalty function. In addition, the penalization functions do not have to be in the same form for association weights and attention strengths. For example, in order to pay attention to a smaller number of feature dimensions it seems more sensible to use M1-4b or M1-5 with small q values for the attention parameters, because the relative but not absolute attention strength values are usually considered. In contrast, either choice seems appropriate for the association weight parameters where raw values are usually used.

Model 2: Bias via asymmetric distribution.

In the present model, random numbers are drawn from an asymmetric distribution with its mode equal to zero. Thus, as in the previous model, the probability of drawing a random number r from the vicinity of current values (i.e., vicinity of zero) is still the highest

$$P(0 - \varepsilon < r < 0 + \varepsilon) > P(M - \varepsilon < r < M + \varepsilon) \qquad \text{(M2-1)}$$

for all $M \neq 0$.

However, unlike the previous model, for a particular parameter value, the probability of drawing a random number which will lead its updated value toward zero is higher than that of a random number that leads to the opposite direction. In other words, when the association weight value, w_{kj} is negative, then the probability of drawing a positive number is greater than a negative number; when the weight is positive, then the opposite is true, or

$$P\left(r^w > 0 \mid w_{kj} > 0\right) < P\left(r^w < 0 \mid w_{kj} > 0\right)$$

$$P\left(r^w > 0 \mid w_{kj} < 0\right) > P\left(r^w < 0 \mid w_{kj} < 0\right) \qquad \text{(M2-2)}$$

For the attention strength parameter α_i the probability of drawing a negative random move is larger than for a positive move, assuming that α_i is constrained to be positive, thus,

$$P\left(r^\alpha > 0\right) < P\left(r^\alpha < 0\right). \qquad \text{(M2-3)}$$

Parameter updates are accomplished by the following functions:

$$w_{kj}^{t+1} = w_{kj}^t + r_{kj}^w \qquad \text{(M2-4)}$$

$$\alpha_i^{t+1} = \alpha_i^t + r_i^\alpha \qquad \text{(M2-5)}$$

where $r_{kj}^w \sim \operatorname{sgn}\left(w_{kj}^t\right) \cdot \Phi^w(\cdot)$ and $r_i^\alpha \sim \Phi^\alpha(\cdot)$

The random movement r_m is drawn from the negatively skewed distributions for α_i and w_{kj} if w_{kj} is positive, and from the positively skewed distributions for negative w_{kj}. Thus, the expected value of the distance of the random movement leading the learnable parameters to zero is greater than that of the opposite direction. This makes the model to decrease values of "irrelevant" parameters quickly.

There are several asymmetric distributions, and the χ^2 (Eq. M2-6, Figure 1, left panel) and Rayleigh (M2-7 & Figure 1, right panel) distributions are examples of asymmetrical distributions.

$$f(x \mid v) = \frac{1}{\Gamma(v/2)}\left(\frac{1}{2}\right)^{v/2} x^{\frac{v}{2}-1} \exp\left(\frac{-x}{2}\right) \qquad \text{(M2-6)}$$

where $\Gamma(\cdot)$ is a gamma function, v is the degree of freedom.

$$f(x \mid b) = \frac{x}{b^2} \exp\left(\frac{-x^2}{2b^2}\right) \qquad \text{(M2-7)}$$

where b is the Rayleigh distribution parameter.

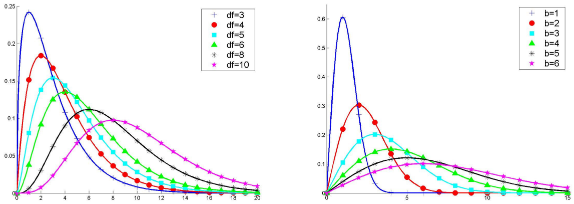

Figure 1. Example asymmetric distributions. Left panel: χ^2 distributions with several different distribution parameters. Right panel: Rayleigh distributions with several different distribution parameters.

Since the modes of these asymmetric non-negative distributions are not zero, and the distribution parameters affect both central tendencies and spreads of the distributions, the random numbers should be transformed as:

$$r = -s^t\left(f(x) - MODE\left(f(x)\right)\right) \qquad \text{(M2-8)}$$

where s^t is a time-dependent scalar controlling the width of the search areas. This ensures that the mode of the transformed random variable is zero and thus satisfies M2-1. Note that the distribution parameter v or b may be selected a priori and held constant throughout the training, or they can be time dependent so that the model starts with a highly skewed distribution and terminates with a near normal distribution, or vice versa.

While the present biased learning model (bias via thorough searches around zero) may be interpreted as active bias, actively trying to reduce the effective number of parameters or simplifying concepts, the bias via regularization (Model 1) may be interpreted as passive bias, involuntary resulting in simpler concept because of the limitation of mental capacity.

Simulations

Here, I examined how the two new biased stochastic learning models account for individual differences in attention learning. To do this, I simulated the results of an empirical study on classification learning, Study 2 of Matsuka (2002). In this study, there were two perfectly redundant feature dimensions, Dimension 1 & Dimension 2 (see Table 1), and those two dimensions are also perfectly correlated with 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 were collected in the empirical study. The measures of attention used were based on feature viewing time, as measured in a MouseLab-type interface (Bettman, Johnson, Luce, & Payne, 1993).

The empirical results that I am trying to simulate indicated that 13 out of 14 subjects were able to categorize the stimuli almost perfectly (Figure 2, top left panel). The aggregated results suggest that on average subjects paid attention to both of the correlated dimensions approximately equally (Figure 3, top middle panel). However, more interestingly when the attention data were analyzed per individual, it was found that many subjects tended to pay attention primarily to only one of the two correlated dimensions, particularly in the late learning blocks as shown in Figure 2, top row third column (Matsuka & Corter, 2003). This suggests that subjects used only the minimal necessary information for this task.

Simulation method: There were three ALCOVE-type models in the present simulation study, namely ALCOVE with stochastic learning (ASL; Matsuka & Corter, 2003a, 2004); ALCOVE with a regularized stochastic learning (ARSL); and ALCOVE with the Rayleigh distribution-based stochastic learning (ARAY). The standard ALCOVE will not be evaluated in the present simulation study, because its standard gradient learning method was shown to be unsuccessful in replicating individual difference when attention allocation is initialized equally (Matsuka & Corter, 2003a, 2004).

All three models were run in a simulated training procedure to learn the correct classification responses for the stimuli of the experiment. ARAY was run for 300 blocks of training, where each block consisted of a complete set of the training instances, while ASL and ARSL were run for 500 training blocks. For each model, the final results are based on 50 replications.

The model configurations (e.g., type of distribution, temperature decreasing rate & function, search ranges) for ASL and ARSL were the same except for the additional parameter-penalization functions incorporated in RSL to model biased processes in category learning. The random numbers for these two models were drawn from the Cauchy distribution, and its random number generation algorithm was based on Ingber (1989). For ARSL, the ridge penalty (Equation M1-3a) was imposed on the association weights, and a subset selection method (M1-4b with $\zeta = 0.1$) was used for the *relative* attention strengths.

For ARAY, a (pseudo) random number generator function from MATLAB Statistical Toolbox (MathWorks, 2001) was used to generate random numbers, and its transforming scalar s (see Eq. M2-8) was exponentially decreased during the learning. For all models, an exponential function was used as the temperature decreasing function. Models' user-definable parameters (e.g., initial temperature, temperature decreasing rate, ζ, and etc…) were selected arbitrarily.

Table 1: Stimulus structure used in Study 2 of Matsuka

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

Results: All three models correctly replicated aggregated or averaged relative attention allocations to the four feature dimensions (Figure 2, second column). However, there are some minor differences in their predictions; ARAY paid less attention to non-diagnostic dimensions than ASL, which in turn paid less attention to those dimensions as compared with ARSL. Qualitatively, ARSL appears to be the most successful in replicating not paying attention to both Dimension 1 and 2 equally, while ASL appears to be least successful in this regard. ARAY was similarly unsuccessful, overestimating the proportion of people who would attend to both of the correlated dimensions equally. A noticeable difference between ARAY and other two models is that ARAY virtually ignored non-diagnostic feature dimensions and paid attention exclusively to either or both Dimensions 1 and 2.

Among all three models, the proportion of sub-zero association weights for ARAY was the largest (Figure 2, fourth column), indicating it yielded simpler category conceptions than the other two models. Here, the notion of simplicity (or complexity) is directly related to numbers of effective (i.e., non-zero, or non-subzero) association weights and attention parameters. When compared with the distribution of the association weights of ASL, the proportion of sub-zero weights for ARSL was larger, indicating penalizing processes incorporated in ARSL resulted in simpler configuration. Note that the model configurations and settings for ASL and ARSL were the same expect for the regularization process incorporated in ARSL. Thus, the straightforward comparison of ASL and ARSL seems reasonable. However, because ARAY and ARSL had different parameter settings, interpreting the comparisons of distributions of the weights for ARAY and ARSL or ASL should be done with care.

In sum, the stochastic learning model with the regularizing processes penalizing mentally-expensive complex category conceptions (i.e. ARSL) appears to be the most successful model capturing human category learning trends that appeared biased, heuristic, and/or less optimal.

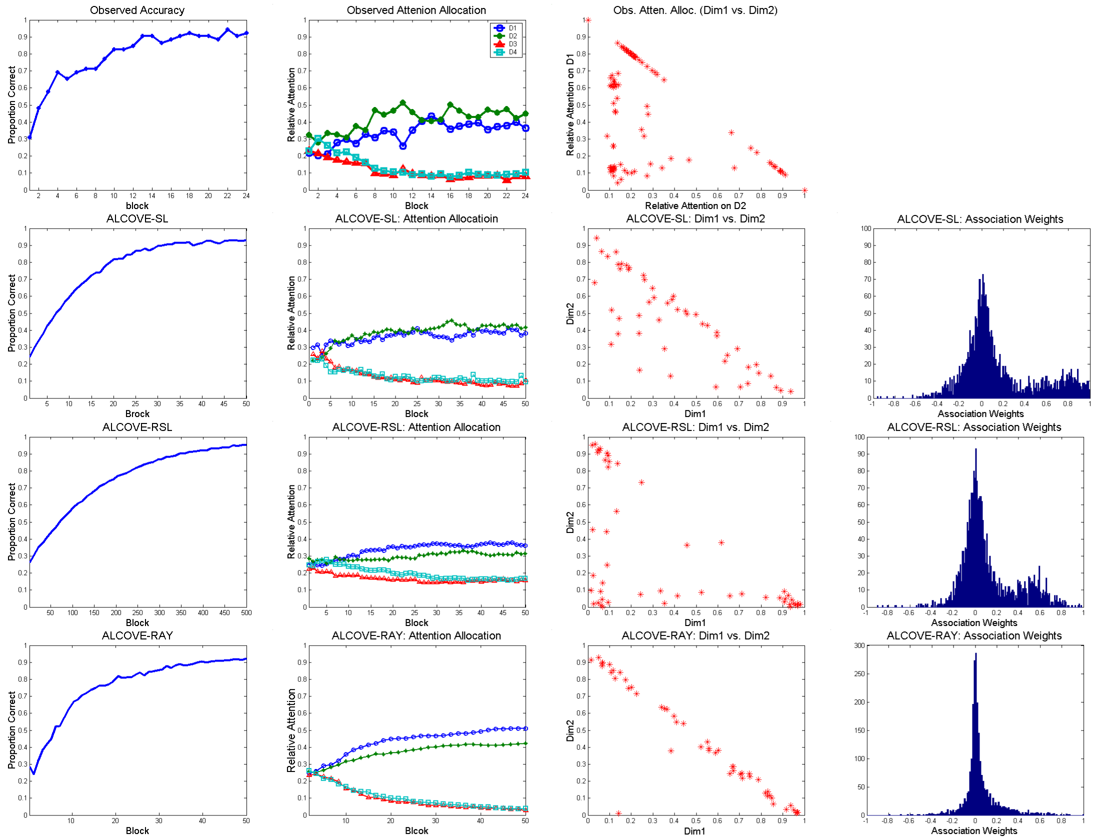

Figure 2. Results of the simulation study. Top row: Observed empirical results of Matsuka & Corter (2003b). The graphs on the first column show observed and predicted classification accuracy, second column shows relative attention allocation for the four feature dimensions; third column compares relative attention allocated to Dimensions 1 and 2 for the last four blocks, where each dot represents an observation. Fourth column shows histograms for the final association weights. Second row shows results of ALCOVE-SL; Third row, ALCOVE-RSL; Fourth Row, ALCOVE-RAY.

Discussion: Although there are 12 unique exemplars in the stimulus set, there are only four exemplars (one from each category) needed for a perfect categorization. Then, one might wonder if people would utilize all the exemplars or not. The distribution of ARAY's association weights may suggest that there are several "dead" or inactive exemplars whose association weights are all zero or near-zero, not being utilized for categorization. This characteristic along with not paying attention to irrelevant feature dimensions may suggest that ARAY replicates learning of an efficient learner, who utilizes a lesser amount of information. In contrast, ARSL predicts that people would utilize more than necessary information. In terms of attention allocation, the empirical results indicate that some people do try utilizing irrelevant information, suggesting that ARSL is more descriptive than other models. This suggests that people may not actively being biased, searching for simpler concepts (i.e., Model 2). Rather it suggests that biases may be caused by the limited mental capacity, involuntarily resulting in simpler concepts.

Discussion

RULEX vs. Stochastic Learning: The stochastic learning's take-all-or–none parameter updating strategy may be considered as a type of hypothesis testing learning model, which makes it similar to the RULEX model (Nosofsky, Palmeri, & McKinley, 1994). 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. Thus, for Matsuka's (2002) stimuli set, RULEX would predict that everyone would allocate his/her attention exclusively to the one of the two diagnostic dimensions. Whereas the stochastic learning would predict some paying attention to either one of the two dimensions, another paying attention to both, and others distributing attention in some other combinations, since, as an exemplar-based model, it can minimize classification error with several different attention allocation patterns (i.e., it can learn to classify stimuli without "optimal" or "rational" attention distribution). 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 parameter (i.e., association weight & attention allocation) configurations, corresponding to possible individual differences. In contrast, RULEX would always predict that people pay attention to the least number of dimensions, which may be a too normative prediction.

Gradient-type vs. Stochastic Learning: For two perfectly redundant feature dimensions, a gradient-type learning algorithm in general would allocate the same amounts of attention to the two dimensions, or its attention learning curves for the two dimensions would be parallel. In contrast, (biased) stochastic learning could result in asymmetric attention allocation to the two dimensions, and its attention learning curves are not necessarily parallel. In these regards, stochastic learning's predictions appear more realistic than those of gradient-type learning. However, this point alone does not necessarily indicate stochastic learning is what people would do. Perhaps, a gradient-type learning with some stochastic elements or errors might, as well, result in more "realistic" predictions.

Conclusion

Biased stochastic learning is a descriptive model of heuristic learning that prefers a simpler conception of categories in which less mental effort seems to be needed. Although the present two stochastic learning algorithms are intended to model such bias, the algorithms appear to be modeling two different types of learners, namely "ordinary people" and "proficients". The simulation study indicates that modeling biased learning via parameter-configuration regularization was the most successful in replicating the empirical results (i.e., ordinary people). In contrast, biased learning via asymmetric distributions appears to be more optimal or rational model, paying attention to only diagnostic feature dimensions and having smaller numbers of effective association weights (proficient-like concepts).

Although the present study supports biased stochastic learning's descriptive validity, more comprehensive simulation studies would be useful in evaluating the present learning models.

References

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.

Cherkassky, V. & Mulier, F. (1997). *Learning from data: Concepts, Theory, and Methods.* New York: Wiley

Hinton, G E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D.E. Rumelhart & J.L. McClelland (Eds.) *Parallel distributed processing: Explorations in microstructure of cognition*. Cambridge, MA: MIT Press.

Ingber, L. (1998). Very fast simulated annealing. *Journal of Mathematical Modelling, 12:* 967-973.

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

Kirkpatrick, S., Gelatt Jr., C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220,* 671-680.

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.

MathWorks. (2001). *MATLAB* [Computer Software]. Natick, MA: Author.

Matsuka, T. (2002). Attention processes in computational models of category learning. Unpublished doctoral dissertation. Columbia University, NY.

Matsuka, T. (2003). Generalized exploratory model of human category learning. Accepted for publication.

Matsuka, T & Corter, J. E. (2003a). Stochastic learning in neural network models of category learning. Proceedings of the 25th Annual Meeting of the Cognitive Science Society.

Matsuka, T. & Corter, J. E. (2003b). Empirical studies on attention processes in category learning. Poster presented at 44th Annual Meeting of the Psychonomic Society. Vancouver, BC, Canada.

Matsuka, T. & Corter, J.E (2004). Stochastic learning algorithm for modeling human category learning. Accepted for publication.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1092.

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

Rehder, B. & Hoffman, A. B. (2003). Eyetracking and selective attention in category learning. Proceedings of the 25th Annual Meeting of the Cognitive Science Society, Boston, 2003.