Psychological Review
1992, Vol. 99, No. 1, 22-44

0033-295X/92/$3.00

ALCOVE: An Exemplar-Based Connectionist Model of Category Learning

John K. Kruschke
Indiana University Bloomington

ALCOVE (attention learning covering map) is a connectionist model of category learning that incorporates an exemplar-based representation (Medin & Schaffer, 1978; Nosofsky, 1986) with error-driven learning (Gluck & Bower, 1988; Rumelhart, Hinton, & Williams, 1986). ALCOVE selectively attends to relevant stimulus dimensions, is sensitive to correlated dimensions, can account for a form of base-rate neglect, does not suffer catastrophic forgetting, and can exhibit 3-stage (U-shaped) learning of high-frequency exceptions to rules, whereas such effects are not easily accounted for by models using other combinations of representation and learning method.

This article describes a connectionist model of category learning called ALCOVE (attention learning covering map). Any model of category learning must address the two issues of what representation underlies category knowledge and how that representation is used in learning. ALCOVE combines the exemplar-based representational assumptions of Nosofsky's (1986) generalized context model (GCM) with the error-driven learning assumptions of Gluck and Bower's (1988a, 1988b) network models. ALCOVE extends the GCM by adding a learning mechanism and extends the network models of Gluck and Bower by allowing continuous dimensions and including explicit dimensional attention learning. ALCOVE can be construed as a combination of exemplar models (e.g., Medin & Schaffer, 1978; Nosofsky, 1986) with network models (Gluck & Bower, 1988a, 1988b), as suggested by Estes (1988; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Hurwitz, 1990). Dimensional attention learning allows ALCOVE to capture human performance where other network models fail (Gluck & Bower, 1988a), and error-driven learning in ALCOVE generates interactions between exemplars that allow it to succeed where other exemplar-based models fail (e.g., Estes et al., 1989; Gluck & Bower, 1988b).

ALCOVE is also closely related to standard back-propagation networks (Rumelhart, Hinton, & Williams, 1986). Although ALCOVE is a feed-forward network that learns by gradient descent on error, it is unlike standard back propagation in its architecture, its behavior, and its goals. Unlike the standard back-propagation network, which was motivated by generalizing neuronlike perceptrons, the architecture of ALCOVE was motivated by a molar-level psychological theory, Nosofsky's (1986) GCM. The psychologically constrained architecture results in behavior that captures the detailed course of human category learning in many situations where standard back propagation fares less well. Unlike many applications of standard back propagation, the goal of ALCOVE is not to discover new (hidden-layer) representations after lengthy training but rather to model the course of learning itself by determining which dimensions of the given representation are most relevant to the task and how strongly to associate exemplars with categories.

The purposes of this article are to introduce the ALCOVE model, demonstrate its application across a variety of category learning tasks, and compare it with other models to highlight its mechanisms. The organization of the article is as follows: First, the ALCOVE model is described in detail; then, its ability to differentially attend to relevant or irrelevant dimensions is demonstrated by applying it to the classic category learning task of Shepard, Hovland, and Jenkins (1961) and to the correlated-dimensions situation studied by Medin, Altom, Edelson, and Freko (1982). Next, the interaction of exemplars during learning is demonstrated by showing that ALCOVE accounts for the apparent base-rate neglect observed by Gluck and Bower (1988a, 1988b) and by Estes et al. (1989) and by showing that ALCOVE learns Medin and Schwanenflugel's (1981) nonlinearly separable categories faster than the linearly separable ones. Afterward, the representation used in ALCOVE is contrasted with that used in standard back propagation, and it is shown that ALCOVE does not suffer the catastrophic retroactive interference seen in standard back propagation (McCloskey & Cohen, 1989; Ratcliff, 1990). Finally, I include a provocative demonstration of ALCOVE's ability to exhibit three-stage learning of rules and exceptions (cf. Rumelhart & McClelland, 1986) and speculate how ALCOVE might interact with a rule-hypothesizing system.

The Model

ALCOVE is a feed-forward connectionist network with three layers of nodes. Its basic computations are a direct implementa-

This article is based on a doctoral dissertation submitted to the University of California at Berkeley. The research was supported in part by Biomedical Research Support Grant RR 7031-25 from the National Institutes of Health.

I thank the members of my dissertation committee, Stuart Dreyfus, Jerry Feldman, Barbara Mellers, Rob Nosofsky, and Steve Palmer. I also thank Steve Palmer for his encouragement and helpfulness as my primary adviser. Rob Nosofsky gets special thanks for sharing with me (unpublished) data and many stimulating conversations and for commenting on earlier versions of this article. Roger Ratcliff and two anonymous reviewers of an earlier version of this article also provided very helpful comments.

Correspondence concerning this article should be addressed to John K. Kruschke, Department of Psychology, Indiana University, Bloomington, Indiana 47405. Electronic mail may be sent to kruschke@ucs.indiana.edu.

tion of Nosofsky's (1986) GCM. Like the GCM, ALCOVE assumes that stimuli can be represented as points in a multidimensional psychological space, as determined by multidimensional scaling (MDS) algorithms (e.g., Shepard, 1957, 1962a, 1962b). Each input node encodes a single psychological dimension, with the activation of the node indicating the value of the stimulus on that dimension. For example, if the first node corresponds to perceived size, and the perceived size of the given stimulus is some scale value ν, then the activation of the first node is ν. The activation of the ith input node is denoted a_i^{in}, and the complete stimulus is denoted by the column vector $a^{in} = (a_1^{in}, a_2^{in}, \cdots)^T$. Figure 1 shows the basic architecture of ALCOVE, illustrating the case of just two input dimensions (in general the model can have any number of input dimensions).

Each input node is gated by a dimensional attention strength, α_i. The attention strength on a dimension reflects the relevance of that dimension for the particular categorization task at hand. Before training begins, the model is initialized with equal attention strengths on all dimensions, and as training proceeds, the model learns to allocate more attention to relevant dimensions and less to irrelevant dimensions. Attention Learning is an important aspect of the model and gives ALCOVE the first two letters of its name. The function of the attention strengths will be described in more detail after the hidden nodes are described.

Each hidden node corresponds to a position in the multidimensional stimulus space. In the simplest version of ALCOVE, there is a hidden node placed at the position of every training exemplar. For example, if the input dimensions are *perceived size* and *perceived brightness*, and one of the training stimuli has scale values of size = ν and brightness = ξ, then there is a hidden node placed at the position (ψ, ξ). In a more complicated version, discussed at the end of the article, hidden nodes are scattered randomly across the space, forming a covering map of the input space. The covering map gives ALCOVE the last four letters of its name. Throughout the body of this article, however, the exemplar-based version is used.

For a given input stimulus, each hidden node is activated according to the psychological similarity of the stimulus to the exemplar at the position of the hidden node. The similarity function is the same as that used in the GCM, which in turn was motivated by Shepard's (1957, 1958, 1987) classic theories of similarity and generalization. Let the position of the jth hidden node be denoted as $(h_{j1}, h_{j2}, \cdots)$, and let the activation of the jth hidden node be denoted as a_j^{hid}. Then

$$a_j^{hid} = \exp\left[-c\left(\sum_i \alpha_i |h_{ji} - \alpha_i^{in}|^r\right)^{q/r}\right], \qquad (1)$$

where c is a positive constant called the *specificity* of the node, where the sum is taken over all input dimensions, and where r and q are constants determining the psychological-distance metric and similarity gradient, respectively. In the applications described in this article, separable psychological dimensions are assumed, so a city-block metric ($r = 1$) with exponential similarity gradient ($q = 1$) is used (Shepard, 1987). For integral dimensions, a Euclidean metric ($r = 2$) could be used (e.g., Nosofsky, 1987; Shepard, 1964).

The pyramids in the middle layer of Figure 1 show the activation profiles of hidden nodes, as determined by Equation 1 with $r = q = 1$. Because the activation indicates the similarity of the input stimulus to the exemplar coded by the hidden node, the activation falls off exponentially with the distance between the hidden node and the input stimulus. The city-block metric implies that the iso-similarity contours are diamond shaped. The specificity constant, c, determines the overall width of the activation profile. Large specificities imply very rapid similarity decrease and hence a narrow activation profile, whereas small specificities correspond to wide profiles. Psychologically, the specificity of a hidden node indicates the overall cognitive discriminability or memorability of the corresponding exemplar. The region of stimulus space that significantly activates a hidden node will be loosely referred to as that node's *receptive field.*

Equation 1 indicates the role of the dimensional attention strengths, α_i. They act as multipliers on the corresponding dimension in computing the distance between the stimulus and the hidden node (cf. Carroll & Chang, 1970). A closely related type of attentional weighting was introduced by Medin and Schaffer (1978) in their context model and generalized into the form shown in Equation 1 by Nosofsky (1984, 1986).

The attention strengths stretch and shrink dimensions of the input space so that stimuli in different categories are better separated and stimuli within categories are better concentrated. Consider a simple case of four stimuli that form the corners of a square in input space, as indicated in Figure 2. If the two left stimuli are mapped to one category (indicated by dots), and the two right stimuli are mapped to another category (indicated by xs), then the separation of the categories can be increased by stretching the horizontal axis, and the proximity within categories can be increased by shrinking the vertical axis. Stretching a dimension can be achieved by increasing its attentional value; shrinking can be achieved by decreasing its attentional value. In ALCOVE, the dimensions most relevant to the category distinction learn larger attention strengths, and the less relevant dimensions learn smaller attention strengths.

Each hidden node is connected to output nodes that correspond to the possible response categories. The connection from the jth hidden node to the kth category node has a connection weight denoted w_{kj}. Because the hidden node is activated only by stimuli in a restricted region of input space near its corre-

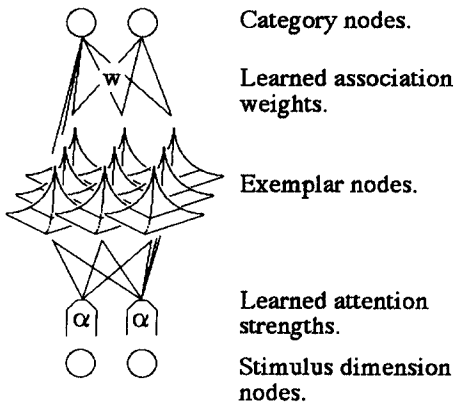

Figure 1. The architecture of ALCOVE (attention learning covering map). (See The Model section.)

Figure 2. Stretching the horizontal axis and shrinking the vertical axis causes exemplars of the two categories (denoted by dots and *x*s) to have greater between-categories dissimilarity and greater within-category similarity. (The attention strengths in the network perform this sort of stretching and shrinking function. From "Attention, Similarity, and the Identification–Categorization Relationship" by R. M. Nosofsky, 1986, *Journal of Experimental Psychology: General, 115,* p. 42. Copyright 1986 by the American Psychological Association. Adapted by permission.)

sponding exemplar, the connection weight is called the *association weight* between the exemplar and the category. The output (category) nodes are activated by the same linear rule used in the GCM and in the network models of Gluck and Bower (1988a, 1988b):

$$a_k^{out} = \sum_{\substack{hid \\ j}} w_{kj} a_j^{hid}. \tag{2}$$

In ALCOVE, unlike the GCM, the association weights are adjusted by an interactive, error-driven learning rule and can take on any real value, including negative values.

To compare model performance with human performance, the category activations must be mapped onto response probabilities. This is done in ALCOVE using the same choice rule as was used in the GCM and network models, which was motivated in those models by the classic works of Luce (1963) and Shepard (1957). Thus,

$$\Pr(K) = \exp(\phi a_K^{out}) / \sum_{\substack{out \\ k}} \exp(\phi a_k^{out}), \tag{3}$$

where ϕ is a real-valued mapping constant. In other words, the probability of classifying the given stimulus into category K is determined by the magnitude of category K's activation (exponentiated) relative to the sum of all category activations (exponentiated).

Here is a summary of how ALCOVE categorizes a given stimulus. Suppose, for example, that the model is applied to the situation illustrated in Figure 2. In this case, there are two psychological dimensions, hence two input nodes; four training exemplars, hence four hidden nodes; and two categories, hence two output nodes. When an exemplar is presented to ALCOVE, the input nodes are activated according to the component dimensional values of the stimulus. Each hidden node is then activated according to the similarity of the stimulus to the exemplar represented by the hidden node, using the attentionally weighted metric of Equation 1. Thus, hidden nodes near the input stimulus are strongly activated, and those farther away in psychological space are less strongly activated. Then the output (category) nodes are activated by summing across all the hidden (exemplar) nodes, weighted by the association weights between the exemplars and categories, as in Equation 2. Finally, response probabilities are computed using Equation 3.

It was stated that the dimensional attention strengths, α_i, and the association weights between exemplars and categories, w_{kj}, are learned. The learning procedure is gradient descent on sum-squared error, as used in standard back propagation (Rumelhart et al., 1986) and in the network models of Gluck and Bower (1988a, 1988b). In the learning situations addressed by ALCOVE, each presentation of a training exemplar is followed by feedback indicating the correct response. The feedback is coded in ALCOVE as *teacher* values, t_k, given to each category node. For a given training exemplar and feedback, the error generated by the model is defined as

$$E = \tfrac{1}{2} \sum_{\substack{out \\ k}} (t_k - a_k^{out})^2, \tag{4a}$$

with the teacher values defined as

$$t_k = \left\{ \begin{array}{l} \max(+1, a_k^{out}) \text{ if the stimulus} \\ \quad \text{is in Category } K, \\ \min(-1, a_k^{out}) \text{ if the stimulus} \\ \quad \text{is not in Category } K. \end{array} \right\}. \tag{4b}$$

These teacher values are defined so that activations "better than necessary" are not counted as errors. Thus, if a given stimulus should be classified as a member of the kth category, then the kth output node should have an activation of at least +1. If the activation is greater than 1, then the difference between the actual activation and +1 is not counted as error. Because these teacher values do not mind being outshone by their students, I call them "humble teachers." The motivation for using humble teacher values is that the feedback given to subjects is nominal, indicating only which category the stimulus belongs to and not the degree of membership. Hence, the teacher used in the model should only require some minimal level of category-node activation and should not require all exemplars ultimately to produce the same activations. Humble teachers are discussed further at the conclusion of the article.

On presentation of a training exemplar to ALCOVE, the association strengths and dimensional attention strengths are changed by a small amount so that the error decreases. Following Rumelhart et al. (1986), they are adjusted proportionally to the (negative of the) error gradient, which leads to the following learning rules (derived in the Appendix):

$$\Delta w_{kj}^{out} = \lambda_w (t_k - a_k^{out}) a_j^{hid}, \tag{5}$$

$$\Delta \alpha_i = -\lambda_\alpha \sum_{\substack{hid \\ j}} \left[\sum_{\substack{out \\ k}} (t_k - a_k^{out}) w_{kj} \right] a_j^{hid} c \left| h_{ji} - a_i^{in} \right|, \tag{6}$$

where the λs are constants of proportionality ($\lambda > 0$) called *learning rates.* The same learning rate, λ_w, applies to all the output weights. Likewise, there is only one learning rate, λ_α, for all the attentional strengths. The dimensional attention strengths are constrained to be nonnegative, as negative values have no psychologically meaningful interpretation. Thus, if Equation 6 were to drive an attention strength to a value less than zero, then the strength is set to zero.

Learning in ALCOVE proceeds as follows: For each presentation of a training exemplar, activation propagates to the category nodes as described previously. Then the teacher values are

presented and compared with the actual category node activations. The association and attention strengths are then adjusted according to Equations 5 and 6. Several aspects of learning in ALCOVE deserve explicit mention.

First, learning is error driven. Both Equations 5 and 6 include the error term $(t_k - a_k^{out})$, so that changes are proportional to error. When there is no error, nothing changes. This is to be contrasted with learning rules that are based on accumulating constant increments on every trial, such as the array-exemplar model (Estes, 1986a, 1986b, 1988; Estes et al., 1989) and context model (Medin & Schaffer, 1978; Nosofsky, 1988b; Nosofsky, Kruschke, & McKinley, in press). In such models, the system changes independently of its actual performance.

Second, because of the similarity-based activations of the hidden nodes, the training exemplars interact during learning. For example, consider two training exemplars that are similar to each other. Because of their similarity, when either one is presented, both corresponding hidden nodes are activated (one just partially); and because learning is proportional to the hidden node activations (see Equation 5), the association strengths from both exemplars are adjusted (as long as there is error present). This interactive property is also to be contrasted with models such as the array-exemplar model, in which learning affects isolated exemplars one at a time (see also Matheus, 1988). The interactive character of learning in ALCOVE is comparable to the competitive nature of learning noted by Gluck and Bower (1988a, 1988b) in their network models and gives ALCOVE the ability to account for the base-rate neglect phenomena they observed, as is described later.

There are other notable implications of interactive learning in ALCOVE. It implies that similar exemplars from the same category should enhance each other's learning. Thus, it suggests that prototypical exemplars should be learned faster than peripheral exemplars, if it can be assumed that prototypical exemplars tend to be centrally located near several other exemplars from the same category. That is desirable insofar as it is also observed in human data (e.g., Rosch, Simpson, & Miller, 1976). Interactive learning also suggests that the shape of the category boundary will have no direct influence on the difficulty of learning the category distinction; rather, difficulty should be based on the clustering of exemplars (subject to the additional complication of attentional learning). In particular, it suggests that it is not necessary for linearly separable categories to be easier to learn than nonlinearly separable categories. Human data again make this a desirable property (Medin & Schwanenflugel, 1981).

A third property of learning in ALCOVE is that attention learning can only adjust the relative importance of the dimensions as given. ALCOVE cannot construct new dimensions to attend to. For example, consider the situation in Figure 3, in which the four training exemplars form the corners of a diamond in the psychological space. Ideally, one might like to stretch the space along the right diagonal to better separate the two categories and shrink along the left diagonal to make within-category exemplars more similar, but ALCOVE cannot do that. Fortunately, it appears that people cannot do that either, as is described later. This anisotropy in attentional learning implies that when modeling human data with ALCOVE, one must be certain that the input dimensions used in the model match the psychological dimensions used by the human subjects.

Figure 3. Attentional learning in ALCOVE (attention learning covering map) cannot stretch or shrink diagonally. (Compare with Figure 2.)

In all the applications described in this article, the psychological dimensions are separable, not integral (Garner, 1974), but the model does not necessarily depend on that. ALCOVE might accommodate psychologically integral dimensions by using a Euclidean distance metric ($r = 2$) in Equation 1 (Nosofsky, 1987; Shepard, 1964). There is evidence to suggest that people can, with effort, differentially attend to psychologically integral dimensions when given opportunity to do so (e.g., Nosofsky, 1987).

In summary, ALCOVE incorporates the exemplar-based representation of Nosofsky's (1987) GCM with error-driven learning as in Gluck and Bower's (1988a, 1988b) network models. ALCOVE extends the GCM in several ways: For learning association weights, it uses an error-driven, interactive rule, instead of a constant-increment rule, that allows association weights in ALCOVE to take on any positive or negative value. ALCOVE also provides a mechanism for attention-strength learning, whereas the GCM has none. ALCOVE extends Gluck and Bower's network models by allowing continuous input dimensions and by having explicit dimensional attention learning. In fitting ALCOVE to human data, there are four free parameters: (a) the fixed specificity c in Equation 1, (b) the probability-mapping constant ϕ in Equation 3, (c) the association weight-learning rate λ_w in Equation 5, and (d) the attention-learning rate λ_α in Equation 6.

Applications

Learning to Attend to Relevant Dimensions

In this section ALCOVE is applied to the category structures used in the classic research of Shepard, Hovland, and Jenkins (1961). There are three reasons for considering the work of Shepard et al.: First, the results of the study provide fundamental human data that any model of category learning should address, and in particular they have served as a benchmark for several recent models (e.g., Anderson, 1991; Gluck & Bower, 1988b; Nosofsky, 1984). Second, the structures described by Shepard et al. are well suited for demonstrating the capabilities of ALCOVE. Third, Shepard et al. argued explicitly that models of categorization based on reinforcement learning and graded generalization could not account for their data unless such models included some (unspecified) mechanism for selective attention. As ALCOVE does include such a mechanism, it faces a direct theoretical and empirical challenge.

The stimuli used by Shepard et al. (1961) varied on three binary dimensions. For example, figures could vary in shape (square vs. triangle), size (large vs. small), and color (filled vs.

open). Each of the resulting eight training exemplars was assigned to one of two categories, such that both categories had four exemplars. It turns out that there are only six structurally distinct types of category assignments. Figure 4 shows the six types, with the eight exemplars indicated by the corners of a cube. The category assignment of an exemplar is indicated by either a filled or blank circle. For example, the top-left cube shows that for Category Type I, Exemplars 1 to 4 are assigned to the *blank* Category, and Exemplars 5 to 8 are assigned to the *filled* category. Any assignment of exemplars to categories, with four exemplars in each category, can be rotated or reflected into one of the structures shown in Figure 4.

A primary concern of Shepard et al. (1961) was to determine the relative difficulty of learning the six category types. Intuitively, Type I should be particularly easy to learn because only information about Dimension 1 is relevant to the categorization decision; variation on Dimensions 2 and 3 leads to no variation in category membership. However, Type II requires attention to both Dimensions 1 and 2 and therefore should be more difficult to learn. (Type II is the exclusive-or [XOR] problem in its two relevant dimensions.) Types III, IV, V, and VI require information about all three dimensions to make correct categorizations, but the dimensions are not equally informative in every type. For example, in Type V, six of eight exemplars can be correctly classified by considering only Dimension 1, with attention to Dimensions 2 and 3 needed only for the remaining two exemplars. On the other hand, Type VI requires equal attention to all the dimensions, because exemplars of each category are symmetrically distributed on the dimensions. (Type VI is the parity problem in three dimensions.) Thus, if it takes more cognitive effort or capacity to consider more dimensions, then Type I should be easiest to learn, followed by Types II, III, IV, V, and VI.

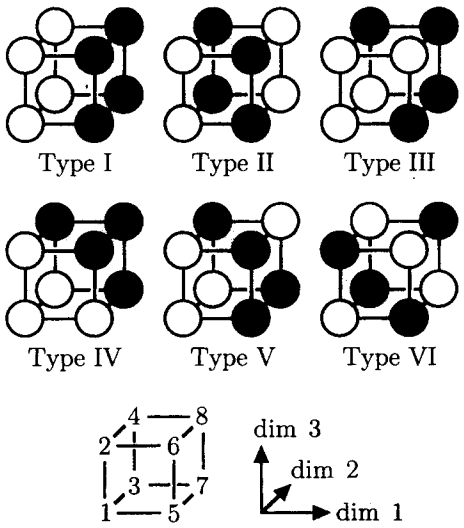

Figure 4. The six category types used by Shepard, Hovland, and Jenkins (1961). (The three binary stimulus dimensions [labeled by the trident at lower right] yield eight training exemplars, numbered at the corners of the lower-left cube. Category assignments are indicated by the open or filled circles. From "Learning and Memorization of Classifications" by R. N. Shepard, C. L. Hovland, & H. M. Jenkins, 1961, *Psychological Monographs, 75,* 13, Whole No. 517, p. 4. In the public domain.)

Shepard et al. (1961) found empirically that the order of difficulty was I < II < (III, IV, V) < VI. That is, Type I was easiest, followed by Type II, followed by Types III, IV, and V (they were very close) and Type VI. Difficulty of learning was measured by the total number of errors made until the subject correctly classified each of the eight exemplars four times in a row. Other measures, such as number of errors in recall, and response time, showed the same ordering.

How does one explain, in a formal quantitative theory, the observed difficulty of the types? Perhaps the most direct approach is a stimulus generalization hypothesis: Category structures that assign highly similar stimuli to the same category and highly dissimilar stimuli to different categories should be relatively easy to learn, whereas structures in which similar stimuli are mapped to different categories and dissimilar stimuli are assigned to the same category should be relatively difficult to learn. Shepard et al. (1961) formalized that hypothesis by measuring interstimulus similarities (inferred from separately obtained identification-confusion data) and by computing the difficulty of category types by considering similarities of all pairs of exemplars from different categories. They considered several variants of the generalization hypothesis, all of which failed to predict the observed order of learning. They argued that "the most serious shortcoming of the generalization theory is that it does not provide for a process of abstraction (or selective attention)." (Shepard et al., 1961, p. 29). The idea was that by devoting attention to only relevant dimensions, confusability of stimuli that differed on those dimensions would be greatly reduced. In that way, Types I and II, especially, would be significantly easier to learn than predicted by a pure generalization theory.

The notion of selective attention was formalized by Nosofsky (1984, 1986) in his GCM. The GCM added attention factors to each dimension of the input space. By using optimal attention weights, which maximized the average percentage correct, or by using attention weights freely estimated to best fit the data, the GCM was able to correctly predict the relative difficulties of the six category types, but the GCM has no attention learning mechanism.

Shepard et al. (1961) considered a variety of learning theories, to see if any provided the necessary attention-learning mechanism. Their answer, in brief, was no. *Cue conditioning* theories, in which associations between single cues (e.g., square) and categories are gradually reinforced, are unable to account for the ability to learn Types II, III, V, and VI, because no single cue is diagnostic of the category assignments. *Pattern conditioning* theories, in which associations between complete configurations of cues (e.g., large, white square) and categories are gradually reinforced, cannot account for the rapidity of learning Types I and II. They concluded

> Thus, although a theory based upon the notions of conditioning and, perhaps, the adaptation of cues at first showed promise of accounting both for stimulus generalization and abstraction, further investigation indicated that it does not, in any of the forms

> yet proposed, yield a prediction of the difficulty of each of our six types of classifications. (Shepard et al., 1961, p. 32)

Gluck and Bower (1988a) combined cue and pattern conditioning into their "configural-cue model." The configural-cue model assumes that stimuli are represented by values on each single dimension, plus pairs of values on each pair of dimensions, plus triplets of values on each triplet of dimensions, and so on. Thus, for the stimuli from the Shepard et al. (1961) study, there are 6 one-value cues (two for each dimension), plus 12 two-value configural cues (four for each pair of dimensions), plus 8 three-value configural cues (the eight full stimuli themselves), yielding a total of 26 configural cues. Each configural cue is represented by an input node in a simple network, connected directly to category nodes. Presence of a configural cue is indicated by activating ($a = +1$) the corresponding input node, and absence is indicated by no activation. The model learns by gradient descent on sum-squared error. For the configural-cue model, Gluck and Bower made no explicit mapping from category-node activations to response probabilities, but in other network models they used the choice function of Equation 3 so that mapping is also assumed here. The configural-cue model has two parameters, the learning rate for the connection weights and the scaling constant ϕ in Equation 3. When applied to the six category types of Shepard et al., the result was that the configural-cue model failed to learn Type II fast enough (see Figure 12 of Gluck & Bower, 1988a), as measured either by cumulative errors during learning or by time until criterion error level is reached. Thus Shepard et al.'s conclusion persists: Some mechanism for selective attention seems to be needed.[1]

ALCOVE was applied to the six category types by using three input nodes (one for each stimulus dimension), eight hidden nodes (one for each training exemplar), and two output nodes (one for each category). It was assumed that the three physical dimensions of the stimuli had corresponding psychological dimensions. In the Shepard et al. experiments, the three physical dimensions were counterbalanced with respect to the abstract dimensions shown in Figure 4; therefore, the input encoding for the simulation gave each dimension equal scales (with alternative values on each dimension separated by one scale unit), and equal initial attentional strengths (set arbitrarily to 1/3). The association weights were initialized at zero, reflecting the notion that before training there should be no associations between any exemplars and particular categories.

In the Shepard et al. (1961) study, the difficulty of any given type was computed by averaging across subjects, each of whom saw a different random sequence of training exemplars. In the simulation, sequence effects were eliminated by executing changes in association weights and attention strengths only after complete epochs of all eight training exemplars. (In the connectionist literature, epoch updating is also referred to as *batch* updating.)

Figure 5 (A and B) show learning curves generated by ALCOVE when there was no attention learning and when there was moderate attention learning, respectively. Each datum shows the probability of selecting the correct category, averaged across the eight exemplars within an epoch. For both graphs, the response mapping constant was set to $\phi = 2.0$, the specificity was fixed at $c = 6.5$, and the learning rate for association weights was $\lambda_w = 0.03$. In Figure 5A, there was no attention learning ($\lambda_\alpha = 0.0$), and it can be seen that Type II is learned much too slowly. In Figure 5B, the attention-learning rate was raised to $\lambda_\alpha = 0.0033$, and consequently Type II was learned second fastest, as observed in human data. Indeed, it can be seen in Figure 5B that the six types were learned in the same order as people, with Type I the fastest, followed by Type II, followed by Types III, IV, and V clustered together, followed by Type VI.

The dimensional attention strengths were redistributed as expected. For Category Type I, the attention strength on the relevant Dimension 1 increased, whereas attention to the two irrelevant Dimensions 2 and 3 dropped nearly to zero. For Type II, attention to the irrelevant Dimension 3 dropped to zero, whereas attention to the two relevant Dimensions 1 and 2 grew (equally for both dimensions). For Types III to VI, all three dimensions retained large attention strengths. Type VI had all of its attention strengths grow, thereby better segregating all the exemplars. Such symmetrical growth of attention is functionally equivalent to increasing the specificities of all the hidden nodes (see Equation 1).

From a model-testing perspective, it is reassuring to note that the range of orderings illustrated in Figure 5 (A and B) are the only orderings that ALCOVE is capable of generating (when $r = q = 1$). When the attention-learning rate is set to higher values, the same ordering as in Figure 5B arises, but with Types I and II learned even faster. When the specificity is made larger (or smaller), the overall separation of the learning curves is lessened (or enlarged, respectively), but the same orderings persist.

---

[1] Recognizing the need to address the dimensional attention issue in the configural-cue model, Gluck and Chow (1989) modified it by making the learning rates on different modules of configural cues self-adaptive. In the case of the Shepard, Hovland, and Jenkins (1961) category types, there were seven different modules of configural cues: a module for each of the three dimensions (each module containing 2 one-value cues), a module for each of the three distinct pairs of dimensions (each module containing 4 two-value configural cues), and a module for the combination of three dimensions (containing 8 three-value configural cues). The learning rates for the seven modules were separately self-modifying according to the heuristic described by Jacobs (1988), which says that if weights change consistently across patterns, then learning rates should increase. The modified configural-cue model was indeed able to capture the correct ordering of the six category types. Did the modified configural-cue model selectively attend to individual dimensions? That is a difficult question to answer. For example, in learning Type II, it seems likely (details were not provided in Gluck & Chow, 1989) that the modified configural-cue model increased its learning rates for the module that combined Dimensions 1 and 2, but decreased the learning rates of all other modules, in particular the modules that individually encode Dimensions 1 and 2. Thus, it increased attention to the combination of dimensions but decreased attention to the individual dimensions. Although this might or might not make sense psychologically, it is clear that further explication of the modified configural-cue model is needed. On the other hand, ALCOVE makes dimensional attention strengths an explicit part of the model and, unlike the modified configural-cue model, allows continuous-valued input dimensions.

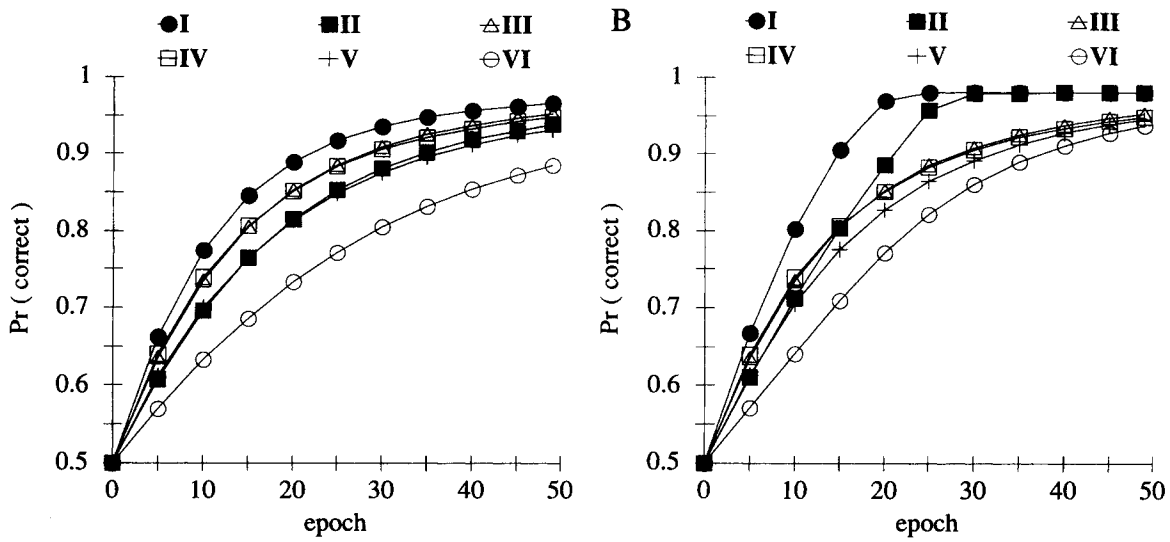

Figure 5. A: Results of applying ALCOVE (attention learning covering map) to the Shepard, Hovland, and Jenkins (1961) category types, with zero attention learning. Here Type II is learned as slowly as Type V (the Type V curve is mostly obscured by the Type II curve). B: Results of applying ALCOVE to the Shepard et al. category types, with moderate attention learning. Note that Type II is now learned second fastest, as observed in human data. Pr = probability.

Adjusting the association-weight learning rate merely changes the overall number of epochs required to reach a certain probability correct.

ALCOVE accounts for the relative difficulty of Shepard et al.'s (1961) six category types by its ability to learn dimensional attention strengths. Such an attentional-learning mechanism is just the sort of thing Shepard et al. called for in their theoretical analyses. It is only fair to note, however, that Shepard et al. also concluded that in addition to abstracting the relevant dimensions, subjects formulated rules for specifying the categories. How ALCOVE might interact with a rule-generating system is discussed in a later section.

Learning to Attend to Correlated Dimensions

Medin et al. (1982) have noted that prototype and other "independent cue" models are not sensitive to correlations between cues. In several experiments, they pitted single-cue diagnosticity against correlated cues to see which would be the better determinant of human categorization performance. They used a simulated medical diagnosis paradigm in which subjects were shown hypothetical patterns of four symptoms. Each of the four symptoms could take on one of two values; for example, watery eyes versus sunken eyes. Subjects were trained on four exemplars of the fictitious disease Terrigitis (T) and four exemplars of the fictitious disease Midosis (M). In this situation the four symptoms are the four dimensions of the stimulus space, and the two diseases are the two alternative categories. The abstract structure of the categories is shown in Table 1. One important aspect of the structure is that the first two symptoms are individually diagnostic, in that p (Terrigitis|Symptom 1 = "1") = .75 and p (Terrigitis|Symptom 2 = "1") = .75, whereas the third and fourth symptoms are not individually diagnostic, each being associated with each disease 50% of the time. Another important aspect of the structure is that the third and fourth symptoms are perfectly correlated in the training set, so that their combination forms a perfect predictor of the disease category. Thus, symptoms three and four are either both 1 or both 0 for cases of Terrigitis, but they are different values for cases of Midosis.

If subjects learn to attend to the correlated third and fourth symptoms to make their diagnoses, then when tested with novel symptom patterns, they should choose Terrigitis whenever the third and fourth symptoms agree. On the other hand, if subjects learn to use the first and second symptoms, then they should choose Terrigitis more often when those symptom values are 1.

Subjects were trained on the first eight exemplars of Table 1 using a free-inspection procedure. Unlike training paradigms in which stimuli are shown sequentially with a definite frequency, Medin et al. (1982) allowed their subjects to freely inspect the eight exemplars during a 10-min period (each exemplar was written on a separate card). After the 10-min training period, subjects were shown each of the possible 16 symptom combinations and asked to diagnose them as either Terrigitis or Midosis. The results are reproduced in Table 1. Three important trends are evident in the data. First, subjects were fairly accurate in classifying the patterns on which they had been trained (Exemplars T1–T4 and M1–M4), choosing the correct disease category 80% of the time. Second, subjects were sensitive to the diagnostic value of the first and second symptoms, in that Novel Patterns N3 and N4 were classified as Terrigitis more often than Patterns N1 and N2, and Patterns N5 and N6 were classified as Terrigitis more often than Patterns N7 and N8. Third, subjects were also apparently quite sensitive to the correlated features, because they classified Patterns N1 to N4,

Table 1
Patterns Used by Medin, Altom, Edelson, and Freko (1982, Experiment 4) and Probabilities of Classifying as Terrigitis After Training

Exemplar	Symptoms	Observed	ALCOVE	Config cue
T1	1 1 1 1	.88	.82	.76
T2	0 1 1 1	.89	.78	.76
T3	1 1 0 0	.73	.82	.76
T4	1 0 0 0	.77	.78	.76
M1	1 0 1 0	.12	.22	.25
M2	0 0 1 0	.17	.18	.25
M3	0 1 0 1	.25	.22	.25
M4	0 0 0 1	.33	.18	.25
N1	0 0 0 0	.53	.59	.44
N2	0 0 1 1	.53	.59	.44
N3	0 1 0 0	.75	.64	.46
N4	1 0 1 1	.67	.64	.46
N5	1 1 1 0	.45	.41	.58
N6	1 1 0 1	.38	.41	.58
N7	0 1 1 0	.36	.36	.55
N8	1 0 0 1	.28	.36	.55

Note. Exemplar Labels T1–T4 refer to the four training exemplars for Terrigitis, and exemplar Labels M1–M4 refer to the four training exemplars for Midosis. Exemplar Labels N1–N8 refer to novel test patterns. Config cue = configural-cue model; ALCOVE = attention learning covering map. Data are from "Correlated Symptoms and Simulated Medical Classification" by D. L. Medin, M. W. Altom, S. M. Edelson, & D. Freko, 1982, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, p. 47. Copyright 1982 by the American Psychological Association. Adapted by permission.

for which Symptoms 3 and 4 agree, as Terrigitis more than 50% of the time, and they classified patterns N5 to N8, for which Symptoms 3 and 4 differ, as Terrigitis less than 50% of the time.

The fourth column of Table 1 shows the results of applying ALCOVE. Eight hidden nodes were used, corresponding to the eight training exemplars. ALCOVE was trained for 50 sweeps, or epochs, on the eight patterns (T1 to T4 and M1 to M4). Association weights and attention strengths were updated after every complete sweep through the eight training patterns, because Medin et al. (1982) did not present subjects with a fixed sequence of stimuli. Best fitting parameter values were $\phi = 0.845$, $\lambda_w = 0.0260$, $c = 2.36$, and $\lambda_\alpha = 0.00965$, yielding a root-mean-squared deviation (RMSD) of 0.104 across the 16 patterns. (The number of epochs used was arbitrary and chosen only because it seemed like a reasonable number of exposures for a 10-min free-inspection period. The best fit for 25 epochs, for example, yielded an RMSD identical to three significant digits.)

All three of the main trends in the data are captured by ALCOVE. The trained exemplars were learned to 80% accuracy. The diagnosticities of the lst two symptoms were picked up, because Patterns N3 and N4 were classified as Terrigitis with higher probability than Patterns N1 and N2, whereas Patterns N5 and N6 were classified as Terrigitis more often than Patterns N7 and N8. It is important to note that the correlated symptoms were detected, because Patterns N1 to N4 were classified as Terrigitis with more than 50% probability, and Patterns N5 to N8, with less than 50% probability.

ALCOVE accounts for the influence of correlated dimensions by increasing attention to those dimensions. When the attention-learning rate is very large, then the correlated Symptoms 3 and 4 get all the attention, and Symptoms 1 and 2 are ignored. On the contrary, when attentional learning is zero, then the diagnosticities of the first two dimensions dominate the results. The results reported in Table 1 are for an intermediate attentional-learning rate, for which Symptoms 3 and 4 get more attention than Symptoms 1 and 2, but some attention remains allocated to Symptoms 1 and 2.

The configural-cue model was also fitted to these data. For this situation, the configural-cue model requires 80 input nodes: 8 singlet nodes, 24 doublet nodes, 32 triplet nodes, and 16 quadruplet nodes. The model was trained for 50 epochs with epoch updating. The best-fitting parameter values were $\phi = 0.554$ and $\lambda_w = 0.0849$, yielding an RMSD of 0.217 across the 16 patterns, more than twice the RMSD of ALCOVE. As is clear from the results shown in Table 1, the configural-cue model is completely unable to detect the correlated symptoms, despite the presence of doublet nodes that are sensitive to pairwise combinations of dimensions. Contrary to human performance, the configural-cue model classifies Patterns N1 to N4 as Terrigitis with less than 50% probability and Patterns N5 to N8 with more than 50% probability. That qualitative reversal is a necessary prediction of the configural-cue model and cannot be rectified by another choice of parameter values.

Gluck, Bower, and Hee (1989) showed that if only single symptoms and pairwise symptom combinations were used, with no three-way or four-way symptom combinations, then correlated symptoms could be properly accentuated (for Experiment 3 from Medin et al., 1982). However, by not including the higher order combinations, the model was told a priori that pairwise combinations would be relevant, which begs the fundamental question at issue here: namely, how it is that the relevance is learned.

The simulation results shown in Table 1 are to be construed qualitatively, despite the fact that they are quantitative best fits. That is because the free-inspection training procedure used by Medin et al. (1982) might very well have produced subtle effects caused by subjects exposing themselves to some stimuli more frequently than to others, or studying different exemplars later in training than early on. The simulations, on the other hand, assumed equal frequency of exposure and constant relative frequencies throughout training. Moreover, there is a disparity in the number of parameters in the two models: ALCOVE has four, whereas the configural-cue model has two. Nevertheless, the qualitative evidence is clear: Because of attention learning, ALCOVE can account for sensitivity to correlated dimensions, whereas the configural-cue model cannot.

Interactive Exemplars and Base-Rate Neglect

The previous sections emphasized the role of attention learning. This and the next section, instead, emphasize the learning of association weights and illustrate how hidden nodes (exemplars) interact during learning because of their similarity-based activations. In particular, it is shown that ALCOVE can quantitatively fit trial-by-trial learning curves and account for the apparent base-rate neglect observed by Gluck and Bower (1988b),

Estes et al. (1989), Shanks (1990), and Nosofsky, Kruschke, and McKinley (in press).

Like the Medin et al. (1982) research, Gluck and Bower (1988b, Experiment 3) had subjects learn to classify lists of four symptoms as one of two fictitious diseases. The base rates of the two diseases were unequal, with one disease occurring 75% of the time, and the other disease, 25% of the time. The diseases were referred to as either the *common* or *rare* disease, respectively (although subjects learned them using fictitious disease names). Symptoms were binary valued, and their alternative values were denoted s1 and s1*, s2 and s2*, and so on for each of the four symptoms. The correspondence of symptoms with diseases was probabilistic, so that on each trial a disease was selected according to the base rates, and then symptoms were selected according to the conditional probabilities in Table 2. The probabilities were designed so that the conditional probability of the rare disease, given only Symptom s1, was 50%. That is, according to Bayes' Theorem, when base rates are properly taken into account, Symptom s1 is completely undiagnostic by itself.

After considerable training, subjects estimated the probability of the diseases given each symptom alone. It turned out that when given Symptom s1 alone, subjects reliably overestimated the probability of the rare disease, apparently not taking full account of the base rates of the diseases.

To explain that apparent base-rate neglect, Gluck and Bower (1988a, 1988b) considered two candidate models of category learning. One was a simple exemplar-based model, in which all training instances were stored in memory along with their assigned categories. To predict categorization probabilities given a single symptom, the memory was scanned for all exemplars that matched on the given symptom, and the response probability for a category was taken as the frequency of matching exemplars assigned to that category, relative to the total frequency of matching exemplars. The simple exemplar-based model predicted that given Symptom s1 alone, the estimated probability of the rare disease should be .5, because exactly half of the training exemplars containing s1 were assigned to the rare disease. This is a special case of Medin and Schaffer's (1978) context model, in which the similarity of nonmatching features is taken to be zero. Nosofsky et al. (in press) described this in more detail, noting that if the context-model similarity parameters are taken to be nonzero, then the exemplar-based model does even worse.

Table 2
Conditional Probabilities of Disease Symptoms in Four Experiments

	Disease	
Symptom	Rare	Common
s1 (s1*)	.6 (.4)	.2 (.8)
s2 (s2*)	.4 (.6)	.3 (.7)
s3 (s3*)	.3 (.7)	.4 (.6)
s4 (s4*)	.2 (.8)	.6 (.4)

Note. The table indicates, for example, that p (s1|rare) = .6. The base rate of the rare disease was .25, and the base rate of the common disease was .75. Parentheses indicate alternative symptom (•) values and corresponding probabilities.

Gluck and Bower (1988a, 1988b) also considered the "double-node" network model (so-called by Estes et al., 1989), illustrated in Figure 6. In this model, each binary-valued stimulus dimension is represented by a pair of input nodes, one node for each of the alternative values on that dimension. When Symptom s1 was present, its node was activated, and the s1* node was deactivated. Each input node was directly connected to output nodes corresponding to the disease categories. The output nodes were linear, and response probabilities for complete (four-symptom) exemplars were computed as in Equation 3. The connection weights were adapted by gradient descent on error. When given just single symptoms, Gluck and Bower (1988b) used the corresponding connection weights to indicate ordinal estimates of disease probabilities. In subsequent work by Estes et al. (1989), quantitative predictions of choice probabilities, given single symptoms, were computed using Equation 3. The latter approach is also taken here.[2]

Unlike the simple exemplar-based model, the double-node model was able to account for the base-rate neglect. As explained by Gluck and Bower (1988b), the error-driven learning mechanism made individual symptom nodes "compete" for the right to activate the output nodes, and in the context of the other training patterns, symptom s1 was a relatively better predictor of the rare disease than the common disease.

Estes et al. (1989) replicated and extended Gluck and Bower's (1988b) study. First, whereas Gluck and Bower were interested in asymptotic behavior after lengthy training, Estes et al. trained subjects on a single sequence of patterns so that trial-by-trial learning curves could be fitted by competing models. Second, whereas Gluck and Bower obtained explicit probability estimates after training, Estes et al. also obtained choice probabilities for each single symptom presented alone.

Estes et al. (1989) compared an exemplar-based model and (a single-node version of) the double-node model in their abilities to fit the trial-by-trial training data and fit the posttraining single-symptom transfer data. In fitting the training data, the Gluck and Bower network model was superior to the simple exemplar model. In fitting the transfer data, the exemplar model could sometimes give better overall fits, but in no case could it predict that p (rare|s1) > .50. In brief, the exemplar models tested by Gluck and Bower (1988b) and by Estes et al. failed to account for the apparent base-rate neglect. ALCOVE is an exemplar-based model, so it faces a direct challenge by these results.

Nosofsky et al. (in press) carried out partial replications and extensions of the experiments reported by Gluck and Bower

[2] Gluck and Bower (1988b) used a network with a single output node, with +1 indicating Disease A and −1 indicating Disease B. Two output nodes are used here because it is formally equivalent to the single node version when just two categories are used, but unlike the single node version it generalizes naturally to situations involving more than two categories. The formal equivalence is easy to demonstrate: Suppose there are two output nodes that always get equal- and opposite-teacher values, so that $a_1^{out} = -a_2^{out}$ at all times. Then Equation 3 can be rewritten as $\Pr(K) = 1/[1 + \exp(-2\phi a_k^{out})]$, the form used by Gluck and Bower. Compare with Footnote 2 of Gluck and Bower (1988b).

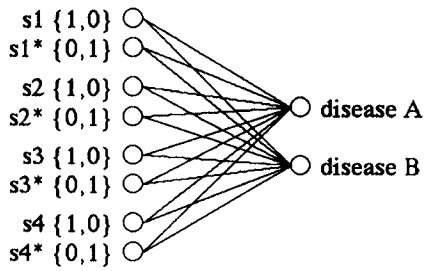

Figure 6. The double node network model of Gluck and Bower (1988b). (Numbers after each symptom indicate the activation of the node pair when that symptom is present. From "From Conditioning to Category Learning: An Adaptive Network Model" by M. A. Gluck and G. H. Bower, 1988, *Journal of Experimental Psychology: General, 117,* p. 239. Copyright 1988 by the American Psychological Association. Adapted by permission.)

(1988a, 1988b) and by Estes et al. (1989). The same sequence of training exemplars and feedback as used by Estes et al. was used in their experiment (hence the same probabilistic structure as shown in Table 2). Instead of using the present versus absent symptoms, as used by Estes et al., Nosofsky et al. (in press) used substitutive symptoms, for example, stuffy nose versus runny nose. One advantage of using substitutive symptoms is that there is no confusion on single-symptom test trials as to whether the unpresented symptoms are completely missing from the stimulus or have the informative value "absent" (Shanks, 1990). The Nosofsky et al. (in press) study also obtained data from a richer set of transfer stimuli, including not only single symptoms but also all pairs, triplets, and complete quadruplets of symptoms and the null pattern. The larger transfer data set is not considered here, as it is fully described in Nosofsky et al. (in press). Instead, only the eight single symptoms considered by Gluck and Bower and by Estes et al. are discussed.

In the Nosofsky et al. (in press) experiment, 84 subjects were trained on the same sequence of 240 exemplars, and then in the transfer stage were presented with patterns without feedback. (Details of the procedure can be found in Nosofsky et al., in press.) The proportion of subjects choosing each category was computed for every trial. The models were fitted to those data, using the sum of squared deviations as the measure of fit.

ALCOVE makes predictions on transfer trials by assuming that missing stimulus dimensions are collapsed. An equivalent method was used by Estes et al. (1989) to test their exemplar model. Functionally, that means that the sum in Equation 1 is taken only over the dimensions actually present in the stimulus. When all dimensions are missing, Equation 1 implies that every hidden node is maximally activated. That allows ALCOVE to predict base rates of the categories by integrating association weights across all the training exemplars.

The models were fitted simultaneously to the training and transfer data, minimizing the sum of the mean squared error on training trials plus the mean squared error on transfer trials. The resulting best fits are shown in Table 3. ALCOVE fits both training and transfer data slightly better than the double-node model. Figures 7 and 8 show the model's predictions for these best simultaneous fits. Figure 8 shows that both models predict that Symptom sl (presented alone) should be classified as the rare disease more than 50% of the time and to about the same degree. (Although both models predict apparent base-rate neglect on Symptom sl, neither fits the transfer results in great detail. Extended versions of the models that address this problem are described in Nosofsky et al., in press.)

The configural-cue model was also fit to the data. Table 3 shows that it did noticeably worse than ALCOVE and the double-node models. In fact, the configural-cue model shows only slight base-rate neglect, with $p(\text{rare}|\text{sl}) = .531$. Because of those inadequacies, the configural-cue model is not shown in Figures 7 and 8.

Some readers might object that these are unfair comparisons because ALCOVE has four free parameters, whereas the double-node (and configural-cue) model has only two. The purpose of the presentation here is to compare the basic versions of the models, and so the inequality in the number of parameters is unavoidable. However, Nosofsky et al. (in press) used versions of the models with equal numbers of parameters. In one set of comparisons, each model was allowed three parameters. For ALCOVE, the attention-learning rate was set to zero, a priori, because the category structure used does not have a strongly asymmetrical distribution of exemplars over dimensions. The double-node model was given a third parameter by including a learning rate on an extra bias node. The bias node was necessary for the double-node model to make predictions about base rates on null patterns, that is, when all dimensions of the stimulus were missing. The results were that ALCOVE consistently did as well as the double-node model, even with equal numbers of parameters.

ALCOVE generates the apparent base-rate neglect on Symptom sl because of interactions between exemplars during learning. For purposes of explanation, consider a simpler case with just two symptoms (two input dimensions). Figure 9 shows the frequencies of rare and common diseases for each combination of Symptoms a and b, out of a total of 104 cases. The top-left cell of Figure 9 indicates that the symptom pair (a, b) occurred 11 times out of 104, with 10 rare cases and 1 common case. The frequencies were selected so that the conditional probability of the rare disease given Symptom a alone (or Symptom b alone) is $15/30 = .50$.

The table in Figure 9 also acts as a geometric representation of the input space. The four cells are the four training exemplars. To model this situation in ALCOVE, there would be four hidden nodes with their receptive fields centered on the four cells of the table. Each hidden node has an association weight with the two disease (category) nodes (not shown).

The node centered on the symptom pair (a, b) should acquire a strong positive association weight with the rare disease node, because (a, b) is the rare disease 10 times as often as it is the common disease. By similar reasoning, one might suppose that the node centered on (a, b*) should acquire a strong negative association weight with the rare disease node, because it is the rare disease only about a third as often as it is the common disease. In fact, when ALCOVE is run on this situation, the magnitude of the negative association weight from (a, b*) is much less than the magnitude of the positive association weight from (a, b). That is because the (a, b*) node has a neighbor, (a*,

Table 3
Fits of ALCOVE, the Double-Node, and Configural-Cue Models to Learning and Transfer Data

Model	RMSD			Parameter value			
	Total	Training	Transfer	ϕ	λ_w	c	λ_α
ALCOVE	.101	.106	.0955	1.06	.0393	2.55	0.0
Double node	.116	.109	.123	1.64	.0122	—	—
Configural cue	.151	.113	.181	2.07	.00312	—	—

Note. Dashes indicate nonapplicability. RMSD = root-mean-squared deviation; ALCOVE = attention learning covering map.

b*), that gains a fairly strong negative association with the rare disease node. The three nodes, (a, b*), (a*, b), and (a*, b*), facilitate each other's learning because of their mutual similarity and because they all tend to be assigned to the common disease, and so their individual association weights remain relatively small. On the other hand, the association weight from (a, b) must become especially large to compensate for its competing neighbors.

When the single symptom (a, –) is presented, both the (a, b) and (a, b*) nodes are fully activated, whereas the (a*, b) and (a*, b*) nodes are both partially (and equally) activated. The net result is that the strong positive association weight from (a, b) to the rare disease node is sufficient to overcome the weaker negative associations from the other exemplars, and the rare disease node receives the greater activation. The model thereby displays apparent base-rate neglect when presented with single symptoms.

In summary, three points have been made in this section. First, the exemplar-based ALCOVE model has been shown to fit the learning and transfer data as well as the double-node model, whereas previously proposed exemplar-based models did not. In particular, ALCOVE accounts for apparent base-rate neglect as well as the double-node model. Second, there is no claim being made that ALCOVE is significantly better than the double-node model in this particular situation. Rather, ALCOVE has an advantage because it also fits several other situations where the double-node model fares less well or is inapplicable. Third, ALCOVE shows apparent base-rate neglect because the combination of error-driven learning and similarity-based hidden-node activations causes exemplars to interact during learning.

Interactive Exemplars in Linearly and Nonlinearly Separable Categories

As suggested in the introduction, ALCOVE is only indirectly sensitive to the shape of category boundaries and is primarily affected by the clustering of exemplars and their distribution over stimulus dimensions. In particular, whether a category boundary is linear or nonlinear should have no direct influence, and it is possible that nonlinearly separable categories would be easier to learn than linearly separable ones.

A case in point comes from the work of Medin and Schwan-

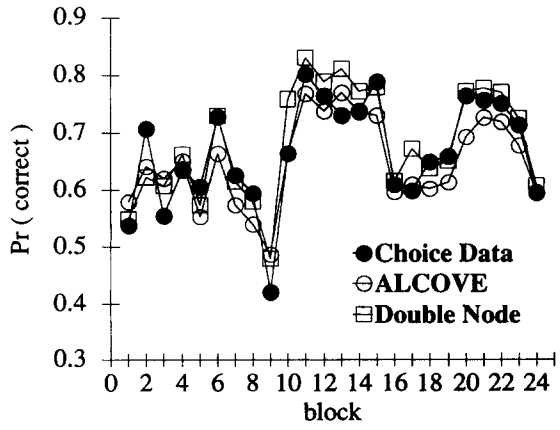

Figure 7. Probability (Pr) of correct category choice during training. (Graph shows means for blocks of 10 trials, although data were fitted trial by trial. Because of averaging within blocks, it appears here that ALCOVE [attention learning covering map] has a worse fit than the double-node model, but the trial-by-trial fit is in fact slightly better. Results shown are for simultaneous fit to training and transfer data.)

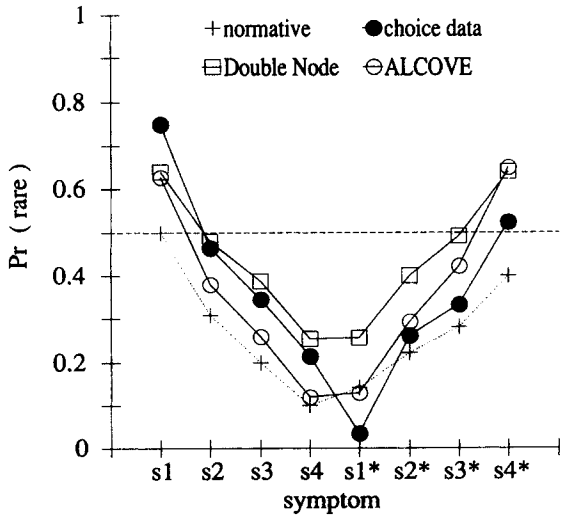

Figure 8. Probability (Pr) of choosing the rare category given single symptoms after training. (Shown are results for simultaneous fit to training and transfer data. Points are connected by lines for visual appeal; no continuum of symptoms is meant to be implied. ALCOVE = attention learning covering map.)

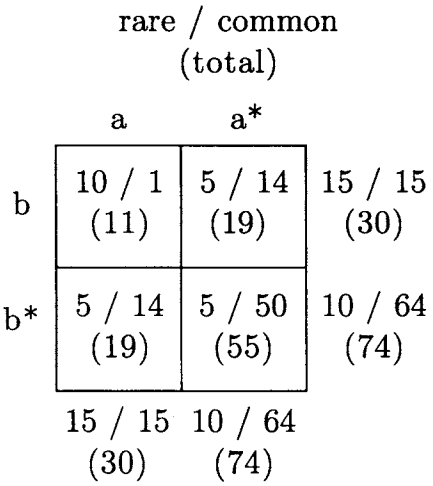

Figure 9. A two-symptom situation to illustrate base-rate neglect in ALCOVE (attention learning covering map.) (Numbers in each cell indicate the frequency that the cell is assigned to the rare or common disease.)

enflugel (1981, Experiment 4). They compared two category structures, shown in Figure 10. One structure was linearly separable, whereas the other was not. The two structures were equalized, however, in terms of mean city-block distance between exemplars within categories and between exemplars from different categories. For example, the mean city-block separation of exemplars within categories for the linearly separable structure is $(2 + 2 + 2 + 2 + 2 + 2)/6 = 2$, and the mean within-category separation for the nonlinearly separable category is the same, $(1 + 2 + 3 + 1 + 2 + 3)/6 = 2$. The mean separation between categories is 1⅔ for both structures.

When human subjects were trained on the two structures, it was found that the linearly separable structure was no easier to learn than the nonlinearly separable structure. This result contradicts predictions of prototype models, such as the single- and double-node models of Gluck and Bower (1988a, 1988b; see Nosofsky, 1991, for a derivation that they are a type of prototype model), but is consistent with models that are sensitive to relational information, such as Medin and Schaffer's (1978) context model, and Nosofsky's GCM. In another experiment run by Medin and Schwanenflugel (1981, Experiment 3), a significant advantage for nonlinearly separable categories was observed.

The configural-cue model is able to show an advantage for the nonlinearly separable category, if the scaling constant ϕ is not too large. Gluck (1991; Gluck et al., 1989) has shown that if the triplet nodes are removed from the configural-cue representation, leaving only the singlet and doublet nodes, the advantage for the nonlinearly separable categories remains. Unfortunately, such a move requires an a priori knowledge of which combinations of dimensions will be useful for the task.

When ALCOVE is applied to these structures, the nonlinearly separable structure is indeed learned faster than the linearly separable structure. This result is true for every combination of parameter values I have tested (a wide range). In particular, attentional learning is not needed to obtain this result. Therefore, it is the interaction of the exemplars, due to similarity and error-driven learning, that is responsible for this performance in ALCOVE. Whereas the mean city-block separations of exemplars were equalized for the two category structures, the mean similarities of exemplars were not equal. ALCOVE exploits that difference in the learning rule for association weights (Equations 1 and 5). The flavor of this explanation is no different from that given for the context model (Medin & Schwanenflugel, 1981). The point is not that ALCOVE necessarily fits these data better than other models with exemplar-similarity-based representations like Medin and Schaffer's (1978) context model but that error-driven learning in ALCOVE does not impair its ability to account for these fundamental data.

Summary

The importance of dimensional attention learning was demonstrated by applying ALCOVE to the six category types from Shepard et al. (1961) and to the Medin et al. (1982) categories involving correlated dimensions. The importance of interaction between exemplars, produced by similarity-based activations and error-driven association-weight learning, was demonstrated in accounting for apparent base-rate neglect and the ability to learn nonlinearly separable categories faster than linearly separable categories. ALCOVE was shown to be quantitatively comparable or superior to the double-node and configural-cue models. Subsequent sections address domains that use continuous dimensions to which the double-node and configural-cue models, as presently formulated, are not applicable.

ALCOVE Versus Standard Back Propagation

As stated in the introduction, ALCOVE differs from standard back propagation in its architecture, behavior, and goals. A *standard back-propagation network* (later referred to as *backprop*) is a feed-forward network with linear-sigmoid nodes in its hidden layer and with hidden weights and output weights that learn by gradient descent on error. Linear-sigmoid nodes have activation determined by

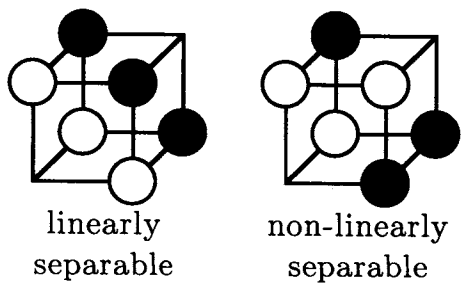

Figure 10. Category structures used by Medin and Schwanenflugel (1981, Experiment 4). (The linearly separable structure is a subset of Type IV in the Shepard, Hovland, and Jenkins, 1961, studies [cf. Figure 4], whereas the nonlinearly separable structure is the corresponding subset from Type III.)

$$a_j^{hid} = 1/[1 + \exp(-\sum_{\substack{in \\ i}} w_{ji}^{hid} a_i^{in})]. \quad (7)$$

The linear-sigmoid function was motivated as a generalized, or smoothed, version of the linear-threshold function in neuron-like perceptrons (Rumelhart et al., 1986). In contrast, the activation functions of ALCOVE were motivated by molar-level psychological theory. The activation profiles of hidden nodes in ALCOVE and in backprop, as determined by Equations 1 and 7, are shown in Figure 11. Three important differences between the activation profiles are evident: First, the hidden node from ALCOVE has a limited receptive field, which means that the node is significantly activated only by inputs near its position. On the contrary, the hidden node from backprop is significantly activated by inputs from an entire half space of the input space. That difference in receptive field size has important consequences for how strongly hidden nodes interact during learning, as is demonstrated shortly. A second difference is that the level contours of the ALCOVE node are iso-distance contours (diamond shaped for a city-block metric), whereas the level contours of the backprop node are linear. (Examples of level contours are shown in Figure 11 by the lines that mark horizontal cross sections through the activation profiles.) This implies that backprop will be especially sensitive to linear boundaries between categories. A third difference between the structure of ALCOVE and backprop is that the linear level contours of the backprop node can be oriented in any direction in input space, whereas attention learning in ALCOVE can only stretch or shrink along the given input dimensions (recall the discussion accompanying Figure 3). Those three differences result in shortcomings of backprop that are now demonstrated with examples.

Insensitivity to Boundary Orientation

When backprop is applied to the six category types of Shepard et al. (1961; see Figure 4), Type IV is learned almost as fast as Type I and much too fast compared to human performance. (Backprop does not learn Type IV quite as quickly as Type I because it is also sensitive to the clustering of exemplars near the boundary; e.g., see Ahmad, 1988.) This result holds over a wide range of learning rates for the two layers of weights, with or without momentum (Rumelhart et al., 1986), for different ranges of initial weight values and over a wide range in the number of hidden nodes. Type IV is learned so quickly by backprop because it can accentuate the diagonal axis through the prototypes of the two categories (Exemplars 1 and 8 in Figure 4), unlike ALCOVE. In other words, the linear level contours of the backprop nodes align with the linear boundary between the categories in Type IV, despite the diagonal orientation of that boundary. ALCOVE cannot direct attention to diagonal axes (see discussion accompanying Figure 3), so it does not learn Type IV so quickly.

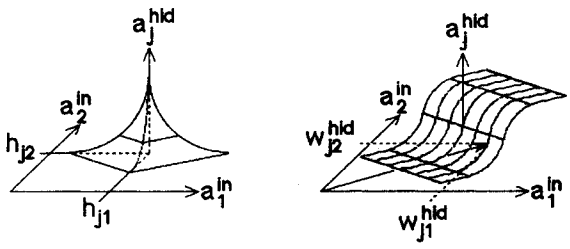

Figure 11. Activation profile of a hidden node in ALCOVE (attention learning covering map) is shown on the left (Equation 1, with $r = q = 1$). (Activation profile of a hidden node in standard back propagation is shown on the right [Equation 7].)

Oversensitivity to Linearity of Boundary

When backprop is applied to the linearly or nonlinearly separable categories of Medin and Schwanenflugel (1981; see Figure 10), the result is that the linearly separable structure is learned much faster than the nonlinearly separable one, contrary to human (and ALCOVE's) performance (e.g., Gluck, 1991). The reason is that the linear level contours of backprop's hidden nodes can align with the linear boundary between categories.

Catastrophic Interference

McCloskey and Cohen (1989) and Ratcliff (1990) have shown that when a backprop network is initially trained on one set of associations, and subsequently trained on a different set of associations, memory for the first set is largely destroyed. Such catastrophic forgetting is not typical of normal humans and is a major shortcoming of backprop as a model of human learning and memory. As ALCOVE is also a feed-forward network that learns by gradient descent on error, it is important to test it for catastrophic forgetting.

A simple demonstration of catastrophic forgetting in backprop is shown in Figure 12 (a–c). The task is to learn the four exemplars in two phases: First learn that $(0, -1) \rightarrow$ "box" and $(-1, 0) \rightarrow$ "circle", then in a second phase learn that $(0, +1) \rightarrow$ "box" and $(+1, 0) \rightarrow$ "circle". The two graphs in panels b and c show typical results of applying backprop and ALCOVE, respectively. Each graph shows probability of correct categorization as a function of training epoch. Phase 1 consisted of Training Epochs 1 to 10, and Phase 2 began after the 10th epoch. Two trends are clear in the backprop results: In Phase 1, generalization performance on the untrained exemplars shifts dramatically to worse than chance, and in Phase 2 performance on the Phase 1 exemplars rapidly decays to worse than chance. On the contrary, ALCOVE shows virtually no interference between Phase 1 and Phase 2 exemplars (Figure 12c).

For the results in Figure 12b, the backprop network was made maximally comparable to ALCOVE. Thus, its input nodes were the same as in ALCOVE, and its output nodes were linear with weights initialized at zero, as in ALCOVE, with probability correct computed with Equation 3. There were 32 hidden nodes, with weights and thresholds initialized to random values between −2.5 and +2.5. Learning rates for output and hidden weights were both 0.06, with epoch updating. The same qualitative trends appear when using other parameter values and numbers of hidden nodes and for standard backprop using linear-sigmoid output nodes and output weights initialized to small random values. The results in Figure 12c were obtained by run-

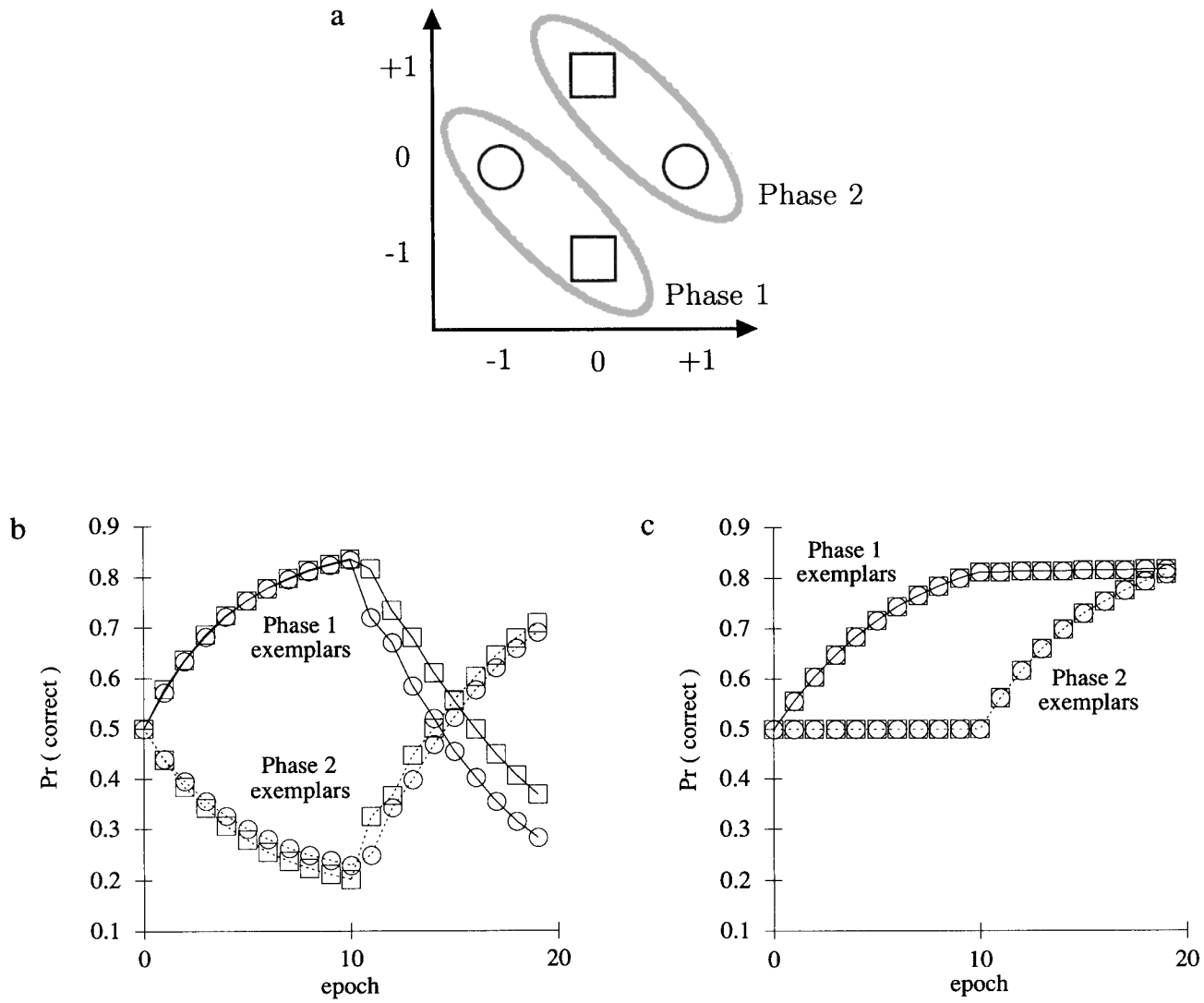

Figure 12. a: Category structure for demonstrating catastrophic forgetting in back propagation and resistance to forgetting in ALCOVE (attention learning covering map.) b: Typical performance of back-propagation on the structure shown in Figure 12a. c: Performance of ALCOVE on the structure shown in Figure 12a. Pr = probability.

ning ALCOVE with four hidden nodes centered on the four exemplars, using $\phi = 1.0$, $\lambda_w = .15$, $c = 2.0$, and $\lambda_\alpha = .06$, with epoch updating.

Backprop shows such severe interference because the receptive fields of its hidden nodes cover such a huge portion of input space. When training on Phase 1, the hidden nodes shift so that their linear level contours tend to align with the right diagonal in Figure 12a so that the two Phase 1 exemplars are accurately discriminated. In addition, nodes that happened to be initially placed in such an opportune orientation have their weights adjusted first and fastest. Unfortunately, those receptive fields cover the untrained Phase 2 exemplars in the same way, and the severe drop in generalization accuracy is the result. When subsequently trained on the Phase 2 exemplars, the same alignment of receptive fields occurs, but the category associations reverse, yielding the reversal of performance on the previously trained exemplars.

The receptive fields of hidden nodes in ALCOVE are much more localized, so that associations from exemplars to categories are not strongly affected by other exemplars, unless the exemplars are very similar. In general, the degree of interference generated in ALCOVE depends on two factors: the size of the receptive fields, as measured by the specificity parameter, c, and whether the exemplars from the two training phases have the same relevant or irrelevant dimensions.

The previous example was used because it was relatively easy to visualize the workings of the two models in terms of how receptive fields get distributed over the stimulus space. The relatively small interference in ALCOVE does not depend on using that particular configuration, however. Similar results

also occur in a situation used by Ratcliff (1990) to demonstrate catastrophic forgetting in backprop. Ratcliff used the "4–4 encoder" problem, in which a network with four input nodes and four output nodes must learn to reproduce isolated activity in each input node on the output nodes. That is, there are just four training patterns: $(+1,-1,-1,-1) \rightarrow (+1,-1,-1,-1)$, $(-1,+1,-1,-1) \rightarrow (-1,+1,-1,-1)$, etc. (These patterns use values of -1 instead of 0 merely to maintain symmetry. Similar qualitative conclusions apply when 0 is used.) The models are initially trained on just the first three training pairs; then, in the second phase of training, they are shown only the fourth pattern pair.

For this demonstration, the backprop network had three hidden nodes, the same number as used by Ratcliff (1990). To maximize comparison with ALCOVE, the four output nodes were linear, and response probabilities were computed with Equation 3, using $\phi = 1.0$. Hidden and output weights had learning rates of 0.2. Hidden weights and biases were initialized randomly in the interval $(-2.5, +2.5)$. Similar qualitative trends obtain for other parameter values, numbers of hidden nodes, etc. (e.g., Ratcliff, 1990); 200 different randomly initialized runs were averaged.

ALCOVE used four hidden nodes, corresponding to the four training exemplars. Specificity of the hidden nodes was set to $c = 2.0$, with association-weight learning rate of 0.05 and attention-learning rate of 0.02. The response scaling constant was set as in the backprop model, $\phi = 1.0$. Similar qualitative trends obtain for other parameter values.

Both models were trained for 100 epochs on the 1st three pattern pairs, then 100 epochs on the fourth pattern pair. Response probabilities at the end of each phase are shown in Table 4. Backprop shows slightly more generalization error in Phase 1, classifying the untrained fourth pattern as one of the three trained patterns more than ALCOVE does. Backprop shows considerable retroactive interference from Phase 2 training: Correct response probabilities on the 1st three patterns drop from 70% to about 40%, and there is considerable bias for backprop to choose the fourth output category even when presented with one of the 1st three input patterns. By contrast, ALCOVE shows no such severe interference. Correct response probabilities on the 1st three patterns decrease only slightly as a consequence of subsequent training on the fourth pattern. The exact amount of interference in ALCOVE is governed by the specificity and the attention-learning rate; the values used here were comparable to those that best fit human learning data in other studies.

In conclusion, the catastrophic forgetting that plagues backprop is not found in ALCOVE because of its localized receptive fields. ALCOVE is able to show significant interference only when the subsequently trained patterns are highly similar to the initially trained patterns or when the second phase of training has different relevant or irrelevant dimensions than the first phase.

Localized Receptive Fields Versus Local Representations

Although the receptive fields of hidden nodes in ALCOVE are relatively localized, the hidden-layer representation is not strictly local, where *local* means that a single hidden node is activated by any one stimulus. In ALCOVE, an input can partially activate many hidden nodes whose receptive fields cover it, so that the representation of the input is indeed distributed over many hidden nodes. (This is a form of continuous coarse coding; see Hinton, McClelland, & Rumelhart, 1986.) However, the character of that distributed representation is quite different from that in backprop because of the difference in receptive fields (Figure 11). One might say that the representation in backprop is more distributed than the representation in ALCOVE and even that the representation in backprop is too distributed.

There are ways to bias the hidden nodes in backprop toward relatively localized representations, if the input patterns are restricted to a convex hypersurface in input space. For example, if the input patterns are normalized, they fall on a hypersphere in input space, in which case the linear level contours of the backprop hidden nodes can "carve off" small pieces of the sphere. For concreteness, consider a two-dimensional input space, so

Table 4
Results of Applying Back Propagation or ALCOVE to the 4–4 Encoder Problem

Input	Back propagation	ALCOVE
	End of Phase 1	
+1 −1 −1 −1	.70 .10 .10 .10	.70 .10 .10 .10
−1 +1 −1 −1	.10 .70 .10 .10	.10 .70 .10 .10
−1 −1 +1 −1	.10 .10 .70 .10	.10 .10 .70 .10
−1 −1 −1 +1[a]	.28 .31 .29 .12	.27 .27 .27 .19
	End of Phase 2	
+1 −1 −1 −1[a]	.40 .07 .08 .45	.69 .09 .09 .13
−1 +1 −1 −1[a]	.08 .39 .07 .46	.09 .69 .09 .13
−1 −1 +1 −1[a]	.08 .07 .40 .45	.09 .09 .69 .13
−1 −1 −1 +1	.10 .10 .10 .70	.10 .10 .10 .70

Note. Data are the probabilities of choosing the corresponding output category. (For $\phi = 1.0$ and four output nodes, asymptotic correct performance in backprop is 0.71.) ALCOVE = attention learning covering map.
[a] Input patterns were not trained during that phase.

that the normalized input patterns fall on a circle. A given linear-sigmoid hidden node "looks down" on this space and makes a linear cut through it, so that all input points to one side of the line produce node activations greater than .5, and all points to the other side of the line produce node activations less than .5. If the linear cut is made near the edge of the circle, then only a small piece of the available input space causes node activations above .5. In particular, Scalettar and Zee (1988) demonstrated that such localized representations are a natural consequence of learning noisy input patterns (with weight decay). Unfortunately, a system that learns a localized representation might also unlearn it, and so it is not clear if the approach taken by Scalettar and Zee could solve the problem of catastrophic forgetting in backprop.

Goals of Backprop Versus Goals of ALCOVE

I have tried to show that backprop and ALCOVE differ in their architecture and behavior. They are also different in their goals. A common goal of applications of backprop is to study the distributed representation discovered by the hidden nodes (e.g., Hanson & Burr, 1990; Lehky & Sejnowski, 1988; Rumelhart et al., 1986; Sejnowski & Rosenberg, 1987) but not to model the course of learning per se. The goals of ALCOVE are quite different. ALCOVE begins with a psychological representation derived from multidimensional scaling that is assumed to remain unchanged during learning. ALCOVE models the course of learning by adjusting attention strengths on the given dimensions and by adjusting association weights between exemplars and categories.

Learning Rules and Exceptions

So far the exemplar-similarity-based representation in ALCOVE has been compared with the featural- and configural-cue representations used in the network models of Gluck and Bower (1988a, 1988b) and with the "half-space receptor" representation in backprop. None of these representations directly addresses the fact that subjects in concept-learning tasks and many categorization tasks consciously generate another representation: rules (e.g., Bourne, 1970; Shepard et al., 1961). Ultimately, the relation of ALCOVE to rule generation must be determined. In this section I outline the beginnings of a theory of how ALCOVE might steer rule generation. The discussion is meant to be exploratory, suggestive, and perhaps provocative, but not conclusive.

One of the most widely known connectionist models of learning is the past-tense acquisition model of Rumelhart and McClelland (1986). That model learned to associate root forms of English verbs with their past-tense forms. The network consisted of input and output layers of nodes that represented *Wickel features,* which are triplets of phoneme features, one feature from each of three consecutive phonemes. The network had no hidden layer, and it learned the connection weights from the inputs to the outputs by using the perceptron convergence procedure, which can be considered to be a limiting case of backprop.

One of the main aspects of past-tense learning that Rumelhart and McClelland (1986) tried to model is the so-called three-stage or U-shaped learning of high-frequency irregular verbs. Children acquire these verbs, such as *go–went,* very early on, in Stage 1. Subsequently, they begin to acquire many regular verbs that form the past tense by adding *ed.* In this second stage, children apparently overgeneralize the rule and regularize the previously well-learned irregular verbs. For example, they might occasionally produce forms like *goed* or *wented.* Finally, in Stage 3, the high-frequency irregular verbs are relearned. Three-stage learning has traditionally been used as evidence that people generate rules. The second stage is explained by suggesting that children literally learn the rule and overapply it. Rumelhart and McClelland's (1986) model had no mechanism for explicit rule generation, so if it could account for three-stage learning, it would pose a challenge to the necessity of rule-based accounts.

The Rumelhart and McClelland (1986) model was indeed able to show three-stage learning of irregular verbs, but that was accomplished only by changing the composition of the training patterns during learning. The network was initially exposed to eight high-frequency irregular verbs and only two regulars. After 10 epochs of training, the network achieved fairly good performance on those verbs. Then the training set was changed to include 334 additional regular verbs and only 76 more irregulars, so the proportion of regulars suddenly jumped from 20% to 80%. As might be expected (especially considering the results on catastrophic forgetting discussed in the previous section), when flooded with regular verbs, the network rapidly learned the regulars but suffered a decrement in performance on the previously learned irregulars. With continued training on the full set of verbs, the network was able to relearn the irregulars. Thus, the transition from Stage 1 to Stage 2 was accomplished only with the help of a deus ex machina, in the form of a radically altered training set. Rumelhart and McClelland defended the approach by saying, "It is generally observed that the early, rather limited vocabulary of young children undergoes an explosive growth at some point in development (Brown, 1973). Thus, the actual transition in a child's vocabulary of verbs would appear quite abrupt on a time-scale of years so that our assumptions about abruptness of onset may not be too far off the mark" (Rumelhart & McClelland, 1986, p. 241). Several critics (e.g., Marcus et al., 1990; Pinker & Prince, 1988) were left unconvinced and argued that a cogent model would have the transition emerge from the learning mechanism, not exclusively from a discontinuity in the training corpus.

Connectionists are left with the challenge of how to model three-stage acquisition of high-frequency irregulars without changing the composition of the training set during learning.[3] It is now shown that ALCOVE can exhibit three-stage learning of

[3] Plunkett and Marchman (1991) showed that a backprop network trained on an unchanging set exhibited micro U-shaped learning, meaning that performance on individual patterns and pattern types fluctuated from epoch to epoch, but gradually improved overall. Their simulations did not exhibit macro U-shaped learning, in which there is a decrease in accuracy on all irregulars over several consecutive epochs, accompanied by an increase in accuracy on regulars, but they argued that such macro U-shaped learning does not occur in children either. Marcus et al. (1990) reported that some aspects of macro U-shaped learning do occur, although they are indeed subtle.

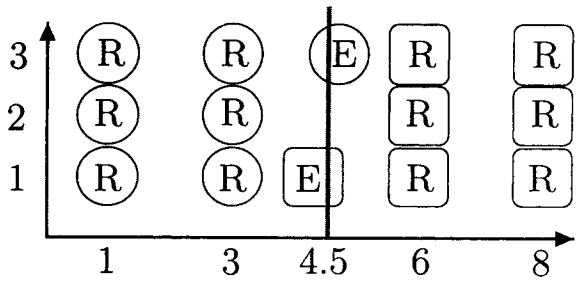

Figure 13. Category structure used for demonstration of three-stage learning of rules and exceptions. (The exemplars marked with an *R* follow the rule, which separates the two categories by the dotted line. Exemplars marked with an *E* are exceptions to the rule. The *x* values of the exceptions were 4.4 and 4.6.)

high-frequency exceptions to rules in a highly simplified abstract analogue of the verb-acquisition situation. For this demonstration, the input stimuli are distributed over two continuously varying dimensions as shown in Figure 13. Of the 14 training exemplars, the 12 marked with an *R* can be correctly classified by the simple rule, "If the value of the exemplar on Dimension 1 is greater than 4.5, then the exemplar is an instance of the box category; otherwise it is in the circle category." This type of rule is referred to as a *Type 1* rule by Shepard et al. (1961), because it segregates members of two categories on the basis of a single dimension. It is also called a *value-on-dimension* rule by Nosofsky, Clark, and Shin (1989), for obvious reasons. In Figure 13 there are two exceptions to the rule, marked with an *E*. The exceptions are presented with higher relative frequency than individual rule exemplars. The analogy to the verb situation is that most of the exemplars are regular, in that they can be classified by the rule, but a few exemplars are irregular exceptions to the rule. The circle and box categories are not supposed to correspond to regular and irregular verbs; rather, they are arbitrary output values (+1 and −1) used only to establish distinct types of mappings on the rule-based and exceptional cases.

ALCOVE was applied to the structure in Figure 13, using 14 hidden nodes and parameter values near the values used to fit the Medin et al. (1982) data: $\phi = 1.00$, $\lambda_\omega = 0.025$, $c = 3.50$, and $\lambda_\alpha = 0.010$. Epoch updating was used, with each rule exemplar occurring once per epoch and each exceptional case occurring four times per epoch, for a total of 20 patterns per epoch. (The same qualitative effects are produced with trial-by-trial updating, with superimposed trial-by-trial "sawteeth," what Plunkett and Marchman, 1991, called micro U-shaped learning.) The results are shown in Figure 14. The learning curve for the exceptions (filled circles) shows a distinct nonmonotonicity so that near Epochs 10 to 15 there is a reversal of learning on the exceptions. (ALCOVE is always performing gradient descent on total error, even when performance on the exceptions drops, because performance on the rule cases improves so rapidly.) The other important feature of the results is that the learning curves for exceptional and rule cases cross over, so that early in training the high-frequency exceptions are learned more accurately, but later in learning the rule cases are learned better. Thus, we have a clear case of three-stage, U-shaped learning.

It should be emphasized that in this demonstration, all parameter values were fixed throughout training, and the composition of the training set was also fixed throughout training. Moreover, there were no order-of-presentation effects because epoch updating was used.

The results shown here should not be construed as a claim that ALCOVE is appropriate for modeling language acquisition. On the contrary, linguistic stimuli, in their natural context, might not be adequately represented by a multidimensional similarity space as demanded by ALCOVE (but cf. Elman, 1989, 1990). Moreover, the results in Figure 14 should not be taken as a necessary prediction of ALCOVE, as some other combinations of parameter values do not show crossover or nonmonotonicities. Rather, the claim is that if such phenomena do occur in human learning, then ALCOVE might very well be able to model those effects.

How does three-stage learning happen in ALCOVE? In the initial epochs, the association weights between exemplars and categories are being established. The association weights from exceptions grow more quickly because the exceptions are presented more frequently. The attention strengths are not affected much in the early epochs because there is not much error propagated back to them by the weak association weights (see Equation 6). Thus, performance on the exceptions is initially better than on the rule cases entirely because of relative frequency.

The second stage begins as the association weights get big enough to back propagate error signals to the attention strengths. Then attention to the rule-irrelevant dimension rapidly decreases (in Figure 13, the vertical dimension shrinks). That has two effects: The rule cases rapidly increase their within-category similarity, thereby improving performance, and the two exceptional cases rapidly increase their between-categories similarity, thereby decreasing accuracy. In other words, once the system learns a little about which exemplars

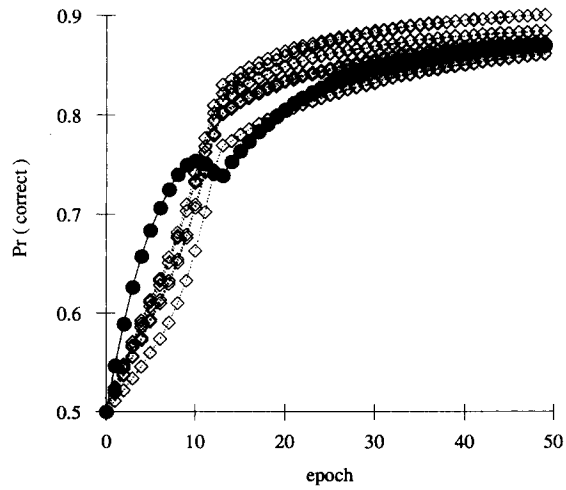

Figure 14. Results of applying ALCOVE (attention learning covering map) to the rules-and-exception structure of Figure 13. (Filled circles show probability of correct classification for exceptions, whereas open diamonds indicate probability of correct classification for the various rule cases. Pr = probability.)

belong in which category, it temporarily ignores the dimension that best distinguishes the exceptions, to benefit the ruly majority.

Such an account of three-stage learning does not prohibit the simultaneous existence of a distinct rule generating system. On the contrary, I believe that a more complete model of human category learning should also include a rule system that would simultaneously try to summarize and generalize the performance of ALCOVE by hypothesizing and testing rules. ALCOVE could help steer the rule-generating system and act as a fallback when adequate rules are not yet found. In such a scenario, the rule-generating system is neither epiphenomenal nor redundant; one major benefit is that rules abstract and unitize category knowledge so it can be transferred to other tasks and stimulus domains.

Perhaps the primary question for such a rule-generating system is which rules should be hypothesized and tested first? The behavior of ALCOVE suggests that one should generate and test rules using the dimensions that are most relevant, where relevance is measured by the dimensional attention strength learned in ALCOVE. This approach is akin to ideas of Bourne et al. (1976), but the notion of driving the rule system with an attention-learning system is new, as far as I know. Details of such an interaction are yet to be worked out; I (Kruschke, 1990a, 1990b) described applications of the idea to the results of Medin, Wattenmaker, and Michalski (1987) and to the learning of exemplars within the six types of Shepard et al. (1961).

In this section I have made two main points: First, ALCOVE is a connectionist network that can show three-stage learning of rules and exceptions without changing the composition of the training set during learning. Second, such a demonstration does not necessarily challenge rule-based accounts; rather, I should like to see future work incorporate ALCOVE-like mechanisms with rule-based systems to capture a wider range of human learning.

Discussion

I have tried to demonstrate that ALCOVE has significant advantages over some other models of category learning. ALCOVE combines an exemplar-based representation with error-driven learning. The exemplar-based representation performs better than other models that also use error-driven learning but with different representations, such as the configural-cue model and backprop. Error-driven learning performs better than other models with exemplar-based representations but different learning rules, such as the array-exemplar model (Estes et al., 1989; Nosofsky et al., in press). In the remainder of the article I discuss variations, extensions, and limitations of ALCOVE.

Placement of Hidden Nodes

All the simulations reported here assumed that a hidden node was placed at the position of each training exemplar, and only at those positions, from the onset of training. That is a reasonable assumption in some circumstances; for example, when the subject previews all the training exemplars (without feedback) before training or when there are so few exemplars that the subject sees them all within a small number of trials. In general, however, the model cannot assume knowledge of the exemplars before it has been exposed to them. There are several ways to deal with that. One way is to recruit new exemplar nodes whenever a novel training exemplar is detected (Hurwitz, 1990). This requires some kind of novelty detection and decision device, which entails the introduction of new parameters, such as a threshold for novelty. An alternative method is to set some a priori bounds on the extent of the input space and randomly cover the space with hidden nodes (Kruschke, 1990a, 1990b). This also entails new parameters, such as the density of the nodes. A third possibility is to recruit a new node for every training trial, regardless of novelty. Careful comparison of these possibilities awaits future research, but I (Kruschke, 1990a, 1990b) reported some preliminary results that the covering map approach fit training data as well as the exemplar approach.

Humble Versus Strict Teacher

The simulations reported here assumed the use of a humble teacher (Equation 4b). This was not an ad hoc assumption, but was motivated by the fact that feedback in category-learning experiments is nominal and does not specify the magnitude of category membership. The humble teachers tell the output nodes that their activation values should reach at least a certain level to indicate minimal membership, but there is no upper limit placed on their activations.

There are situations where a strict teacher is appropriate. Perhaps the most important use of a strict teacher has been the modeling of overexpectation error in animal learning (e.g., Kamin, 1969; Kremer, 1978; Rescorla & Wagner, 1972). Overexpectation occurs when an animal is first trained to associate Conditioned Stimulus (CS) 1 with an unconditioned stimulus (US), denoted $CS_1 \rightarrow US$, then trained on $CS_2 \rightarrow US$, and finally trained on the compound stimulus $(CS_1 + CS_2) \rightarrow US$. The result is that the final training on the compound stimulus actually reduces the individual association strengths from CS_1 and CS_2. A strict teacher with error-driven learning (the Rescorla-Wagner learning rule) can account for that, because at the beginning of training with the compound stimulus $(CS_1 + CS_2)$, the double-strength association overshoots the teacher and is counted as an overexpectation error, causing the individual associations to be reduced. In that situation, however, there is reason to believe that the feedback is encoded by the animal as having a certain magnitude, and not just nominally. For example, in many experiments the feedback was magnitude of electric shock or amount of food.

One difference between humble and strict teachers regards asymptotic performance. Strict teachers demand that all exemplars are equally good members of the category, in that they all activate the category nodes to the same degree. Humble teachers allow more typical exemplars to activate their category nodes more than peripheral exemplars, even after asymptotic training. That difference is robust when measured in terms of category node activations; however, when transformed into response probabilities by the choice rule (Equation 3), the difference is compressed by ceiling and floor effects and becomes very subtle. For the applications reported in this article, the

difference in fits, using humble or strict teachers, is slight. Thus, although I believe the distinction between humble and strict teachers is conceptually well motivated, it remains for future research to decide conclusively which is best for modeling category learning.

Extensions of ALCOVE

Several reasonable extensions of ALCOVE that might allow it to fit a wider array of category learning phenomena, without violating the motivating principles of the model, are possible.

The choice rule in Equation 3 was used primarily because of historical precedents, but it is not a central feature of the model, and there might be better ways of mapping network behavior to human performance. For example, one might instead incorporate random noise into the activation values of the nodes and use a deterministic choice rule such as selecting the category with the largest activation (cf. McClelland, 1991). Also, the particular choice of teacher values in Equation 4b was arbitrary and motivated primarily by the precedent of Gluck and Bower (1988a, 1988b). It might be that a different choice of teacher values, for example, +1 for "in" and 0 (instead of −1) for "not in" would be more appropriate, especially in conjunction with different response rules.

Many researchers have suggested that training has local or regional attentional effects, rather than (or in addition to) global effects (e.g., Aha & Goldstone, 1990; Aha & McNulty, 1989; Medin & Edelson, 1988; Nosofsky, 1988a). ALCOVE is easily altered to incorporate local attention strengths by giving each hidden node j a full set of dimensional attention strengths α_{ji}. In this particular variation there are no new parameters added because there is still just one attention-learning rate. It remains to be seen if exemplar-specific attention strengths, or some combination of exemplar-specific and global attention strengths, can account for an even wider range of data.

A related approach to introducing local attentional effects is to adapt individual hidden node specificities. Specificity learning (by gradient descent on error) would adjust the receptive-field size of individual hidden nodes, so that nodes surrounded by exemplars assigned to the same category would enlarge their receptive fields to encompass those other exemplars, whereas nodes near exemplars assigned to other categories would reduce their receptive fields to exclude those other exemplars. One implication is that asymmetric similarities (Rosch, 1975; Tversky, 1977) would evolve: Peripheral or boundary exemplars would be more similar to central or typical exemplars than vice versa, because the receptive field of the central exemplar would cover the peripheral exemplar, but the receptive field of the peripheral exemplar would not cover the central exemplar.

Another possible extension retains global dimensional attention strengths but changes the dynamics of attention learning. In this article it was assumed that the attention strengths α_i were primitives in the formalization, in that attention strengths were not themselves a function of some other underlying variables. If, however, each attention strength α_i is some nonlinear function of an underlying variable β_i, then gradient descent with respect to β_i will lead to different changes in α_i than gradient descent with respect to α_i itself. For example, suppose we let $\alpha_i = 1/(1 + e^{-\beta_i})$. This has three potentially desirable features: First, it automatically keeps the attention strengths α_i nonnegative, so that it is not necessary to clip them at zero. Second, it automatically keeps the attention strengths bounded above, so that there is a built-in "capacity" limit (cf. Nosofsky, 1986). Third, and perhaps most important, the gradient-descent learning rule for β_i is the same as the learning rule for α_i (Equation 6) except for the inclusion of a new factor, $\partial\alpha_i/\partial\beta_i = \alpha_i(1 - \alpha_i)$. This implies that the attention strength will not change very rapidly if it is near one of its extreme values of +1 or 0. In particular, if the system has learned that one dimension is highly relevant (α_1, nearly 1) and a second dimension is irrelevant (α_2, nearly 0), then it will be reluctant to change those attention strengths. Such an extension might allow ALCOVE to model the ease shown by adults to learn intradimensional feedback reversals relative to interdimensional relevance shifts (Kendler & Kendler, 1962), which ALCOVE cannot capture in its present form (W. Maki, personal communication, October 1990).[4]

Limitations of ALCOVE

ALCOVE applies only to situations for which the stimuli can be appropriately represented as points in a multidimensional psychological similarity space. Moreover, ALCOVE assumes that the basis dimensions remain unchanged during category learning, and it does not apply to situations in which subjects generate new dimensions of representation, or otherwise recode the stimuli, during learning. Predictions made by ALCOVE are therefore based on two sets of premises: One set regards the representational assumptions just stated. The other set regards the exemplar-similarity-based architecture and error-driven learning rules of the model. If ALCOVE should fail to capture data from a given situation, either or both of the sets of premises might be wrong.

Another, perhaps more severe, limitation of ALCOVE is that it does not have a mechanism for hypothesizing and testing rules, whereas people clearly do. As suggested in a previous section, ALCOVE might subserve a rule-generating system, steering its selection of candidate rules. Until such a combination of systems is created, Holland, Holyoak, Nisbett, and Thagard's (1986) assessment of the Rescorla-Wagner learning rule

---

[4] Hurwitz (1990; "hidden pattern unit model Version 2") independently developed a closely related model that had hidden nodes with activation function determined by a multiplicative similarity rule (Medin & Schaffer, 1978). For direct comparison with ALCOVE's hidden nodes, Hurwitz's activation function can be formally reexpressed as follows:

$$a_j^{hid} = \prod_i (1/1 + e^{(\beta_i - k)})^{|h_{ji} - a_i^{in}|} = \exp[-\sum_i \underbrace{\ln(1 + e^{(\beta_i - k)})}_{\alpha_i} |h_{ji} - a_i^{in}|],$$

where k is a constant. Thus, Hurwitz's model can be construed as a version of ALCOVE with $r = q = 1$ in Equation 1 and with $\alpha_i = \ln(1 + e^{(\beta_i - k)})$. Hurwitz's model therefore keeps the attention strengths α_i nonnegative but unbounded above. Gradient descent with respect to β_i results in the right-hand side of Equation 6 except for the absence of the specificity c and the inclusion of a new factor, $\partial\alpha_i/\partial\beta_i = (1 - e^{-\alpha_i})$. That causes attention strengths near zero to be reluctant to change, but causes large attention strengths to change rapidly.

might also apply to ALCOVE: "The limits of [Rescorla and Wagner's] approach can be characterized quite simply—their equation is generally able to account for phenomena that primarily depend on strength revision but is generally unable to account for phenomena that depend on rule generation" (p. 167).

References

Aha, D. W., & Goldstone, R. (1990). Learning attribute relevance in context in instance-based learning algorithms. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 141–148). Hillsdale, NJ: Erlbaum.

Aha, D. W., & McNulty, D. M. (1989). Learning relative attribute weights for instance-based concept descriptions. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 530–537). Hillsdale, NJ: Erlbaum.

Ahmad, S. (1988). *A study of scaling and generalization in neural networks* (Tech. Rep. No. UIUCDCS-R-88-1454). Urbana-Champaign: University of Illinois at Urbana-Champaign, Computer Science Department.

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

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review, 77,* 546–556.

Bourne, L. E., Ekstrand, B. R., Lovallo, W. R., Kellogg, R. T., Hiew, C. C., & Yaroush, R. A. (1976). Frequency analysis of attribute identification. *Journal of Experimental Psychology: General, 105,* 294–312.

Brown, R. (1973). *A first language.* Cambridge, MA: Harvard University Press.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35,* 283–319.

Elman, J. L. (1989). *Representation and structure in connectionist models* (Tech. Rep. No. 8903). San Diego: University of California at San Diego, Center for Research in Language.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General, 115,* 155–174.

Estes, W. K. (1988). Toward a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language, 27,* 196–212.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage–retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556–576.

Garner, W. R. (1974). *The processing of information and structure.* Hillsdale, NJ: Erlbaum.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science, 2,* 50–55.

Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195.

Gluck, M. A., & Bower, G. H. (1988b). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Gluck, M. A., Bower, G. H., & Hee, M. R. (1989). A configural-cue network model of animal and human associative learning. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Gluck, M. A., & Chow, W. (1989). *Dynamic stimulus-specific learning rates and the representation of dimensionalized stimulus structures.* Unpublished manuscript.

Hanson, S. J., & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences, 13,* 471–489.

Hinton, G. E., McClelland, & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (chapter 3). Cambridge, MA: MIT Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction.* Cambridge, MA: MIT Press.

Hurwitz, J. B. (1990). *A hidden-pattern unit network model of category learning.* Unpublished doctoral dissertation, Harvard University.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks, 1,* 295–307.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment.* New York: Appleton-Century-Crofts.

Kendler, H. H., & Kendler, T. S. (1962). Vertical and horizontal processes in problem solving. *Psychological Review, 69,* 1–16.

Kremer, E. F. (1978). The Rescorla-Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 4,* 22–36.

Kruschke, J. K. (1990a). *A connectionist model of category learning.* Doctoral dissertation, University of California at Berkeley. University Microfilms International.

Kruschke, J. K. (1990b). *ALCOVE: A connectionist model of category learning* (Cognitive Science Research Rep. No. 19). Bloomington: Indiana University.

Lehky, S. R., & Sejnowski, T. J. (1988). Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature, 333,* 452–454.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Marcus, G. F., Ullman, M., Pinker, S., Hollander, M., Rosen, T. J., & Xu, F. (1990). *Overregularization* (Occasional Paper No. 41). Cambridge, MA: MIT, Center for Cognitive Science.

Matheus, C. J. (1988). Exemplar versus prototype network models for concept representation (abstract). *Neural Networks, 1*(Suppl. 1), 199.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1–44.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–165). San Diego, CA: Academic Press.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117,* 68–85.

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

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355–368.

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science, 11,* 299–339.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108.

Nosofsky, R. M. (1988a). On exemplar-based exemplar representations: Reply to Ennis (1988). *Journal of Experimental Psychology: General, 117,* 412–414.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Nosofsky, R. M. (1991). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of W. K. Estes.* Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282–304.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. (in press). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel, distributed processing model of language acquisition. *Cognition, 28,* 73–193.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition, 38,* 43–102.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 2,* 285–308.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory.* New York: Appleton-Century-Crofts.

Robinson, A. J., Niranjan, M., & Fallside, F. (1988). *Generalising the nodes of the error propagation network* (Tech. Rep. No. CUED/F-INFENG/TR.25). Cambridge, England: Cambridge University Engineering Department.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7,* 532–547.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 491–502.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by back-propagating errors. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, chapter 8). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 2, chapter 18). Cambridge, MA: MIT Press.

Scalettar, R., & Zee, A. (1988). Emergence of grandmother memory in feed forward networks: Learning with noise and forgetfulness. In D. Waltz & J. A. Feldman (Eds.), *Connectionist models and their implications: Reading from cognitive science* (pp. 309–327). Norwood, NJ: Ablex.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145–168.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42A,* 209–237.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325–345.

Shepard, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review, 65,* 242–256.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27,* 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika, 27,* 219–246.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54–87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13, Whole No. 517).

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Appendix

Derivation of Learning Rules

Here are derived the learning rules used in ALCOVE. Learning of any parameter in the model is done by gradient descent on a cost function such as sum-squared error. The purpose is to determine gradient-descent learning equations for the attention strengths, α_i, and the association weights, w_{kj}^{out}. All the derivations are simple insofar as they involve only the chain rule and algebra. On the other hand, they are complicated insofar as they involve several subscripts simultaneously, and care must be taken to keep them explicit and consistent. Subscripts denoting variables are in lowercase letters. Subscripts denoting constants are in uppercase letters. Vector notation is used throughout the derivations: Boldface variables denote vectors. For example, $\mathbf{a}^{out} = [\cdots a_k^{out} \cdots]^T$ is the column vector of output activation values for the current stimulus.

The General Case

I first compute derivatives using an unspecified cost function C and then treat the specific case of sum-squared error. Suppose that C is some function of the output of the network and perhaps of some other constants (such as teacher values for the output nodes). In general, any parameter x is adjusted by gradient descent on C, which means that the change in x is proportional to the negative of the derivative: $\Delta x = -\lambda_x \partial C/\partial x$, where λ_x is a (nonnegative) constant of proportionality, called the learning rate of parameter x.

I begin by rewriting Equation 1 in two parts, introducing the notation net_j^{hid}:

$$\text{net}_j^{hid} = \Big(\sum_{\substack{in \\ i}} \alpha_i |h_{ji} - a_i^{in}|^r\Big)^{1/r} \text{ and}$$

$$a_j^{hid} = \exp\,[-c(\text{net}_j^{hid})^q], \quad \text{(A1)}$$

where r and q are positive numbers. The special case of $r = 1$ (city-block metric) and $q = 1$ (exponential-similarity decay) are subsequently treated.

Because the output nodes are linear (Equation 2), the derivative of C with respect to the association weights between hidden and output nodes is

$$\frac{\partial C}{\partial w_{KJ}^{out}} = \frac{\partial C}{\partial a_K^{out}} \frac{\partial a_K^{out}}{\partial w_{KJ}^{out}} = \frac{\partial C}{\partial a_K^{out}} a_J^{hid}. \quad \text{(A2)}$$

The derivative $\partial C/\partial a_K^{out}$ must be computed directly from the definition of C, but it can presumably be evaluated locally in the Kth output node. Hence, the weight change resulting from gradient descent is locally computable.

In the applications reported in this article, there was never a need to alter the hidden node positions or specificities. Therefore, I do not compute the derivatives of the hidden node coordinates or specificities, although they certainly can be computed (e.g., Robinson, Niranjan, & Fallside, 1988). Now consider the attention strengths α_I. First note that

$$\frac{\partial C}{\partial \alpha_I} = \frac{\partial C}{\partial \mathbf{a}^{out}} \frac{\partial \mathbf{a}^{out}}{\partial \mathbf{a}^{hid}} \frac{\partial \mathbf{a}^{hid}}{\partial \alpha_I} = [\cdots \partial C/\partial a_{k^{out}} \cdots] \times \begin{bmatrix} \vdots \\ \cdots w_{kj}^{out} \cdots \\ \vdots \end{bmatrix} \begin{bmatrix} \vdots \\ \partial a_j^{hid}/\partial \alpha_I \\ \vdots \end{bmatrix}. \quad \text{(A3a)}$$

Computation of $\partial a_j^{hid}/\partial \alpha_I$ requires a bit more work:

$$\frac{\partial a_j^{hid}}{\partial \alpha_I} = \frac{\partial a_j^{hid}}{\partial \text{net}_j^{hid}} \frac{\partial \text{net}_j^{hid}}{\partial \alpha_I} = -a_j^{hid} cq\,(\text{net}_j^{hid})^{(q-1)} \times \frac{1}{r}\Big(\sum_{\substack{in \\ i}} \alpha_i |h_{ji} - a_i^{in}|^r\Big)^{(1/r-1)} |h_{jI} - a_I^{in}|^r$$

$$= -a_j^{hid} c \frac{q}{r} (\text{net}_j^{hid})^{(q-r)} |h_{jI} - a_I^{in}|^r. \quad \text{(A3b)}$$

Substituting Equation A3b into Equation A3a yields

$$\frac{\partial C}{\partial \alpha_I} = \sum_{\substack{hid \\ j}} \left(\sum_{\substack{out \\ k}} \frac{\partial C}{\partial a_k^{out}} w_{kj}^{out} \right) \times a_j^{hid} c \frac{q}{r} (\text{net}_j^{hid})^{(q-r)} |h_{jI} - a_I^{in}|^r. \quad \text{(A4)}$$

The factors of Equation A4 are all available to input node I if one permits backwards connections from hidden nodes to input nodes that have connection weight equal to the fixed value 1. (Usually in back propagation the backward links are conceived as having the same value as adaptive forward links.) The mechanism for computing the derivatives is the same, in spirit, as that used in "standard" back propagation (Rumelhart, Hinton, & Williams, 1986): The partial derivatives computed at each layer are propagated backwards through the network to previous layers.

Equation A4 reveals some interesting behavior for the adaptation of the attentional parameter α_I. The equation contains the factor

$$F_j = a_j^{hid} (\text{net}_j^{hid})^{(q-r)} |h_{jI} - a_{Iin}|^r.$$

All the other factors of Equation A4 can be considered constants in the present context, so that the change in attention α_I is proportional to F_j. The question now is when is the change large, that is, when is F_j significantly nonzero? The precise answer depends on the values of q and r, but a qualitative generalization can be made about the form of F_j. The graph of F_j (not shown) is a "hyper dumbbell" shape that is centered on the input stimulus, with its axis of symmetry along the Ith dimension. Hence, the attentional parameter is only affected by hidden nodes within the hyper dumbbell region.

Sum-Squared Error

Now consider a specific case for the objective function C, the sum-squared error, as in Equation 4a. Note first that

(*Appendix continues on next page*)

$$\frac{\partial E}{\partial a_K^{out}} = -(t_K - a_K^{out}). \quad \text{(A5)}$$

This derivative (Equation A5) is continuous and well behaved, even with the humble-teacher values. Then Equation A5 can be substituted into each of Equations A2 and A4.

Special Case of $q = r = 1$

In the special case when $q = r$ (and in particular when $q = r = 1$), the learning equation for attention strengths simplifies considerably. In this special case, the term $\frac{q}{r}(\text{net}_j^{hid})^{(q-r)}$ in Equation A4 reduces to 1. The initial computation of a_j^{hid} also simplifies (cf. Equation A1). The learning rules reported in the text (Equations 5 and 6) are the result.

Received June 18, 1990
Revision received February 27, 1991
Accepted April 17, 1991 ■

Correction to Kornblum et al.

In the article "Dimensional Overlap: Cognitive Basis for Stimulus–Response Compatibility —A Model and Taxonomy," by Sylvan Kornblum, Thierry Hasbroucq, and Allen Osman (*Psychological Review*, 1990, Vol. 97, No. 3, pp. 253–270), erroneous data were included in Figure 2. Below are the corrected figure and the original caption.

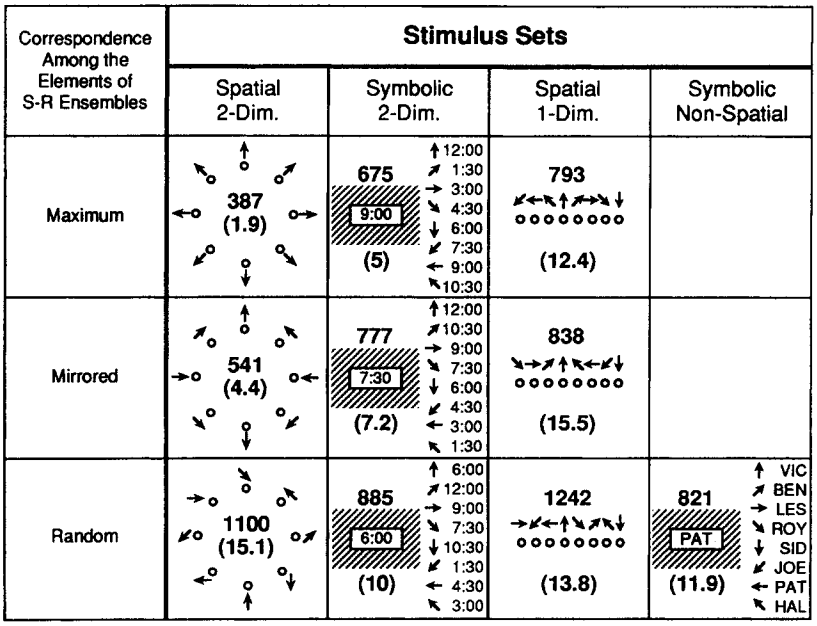

Figure 2. Reaction time and errors (in parentheses) for four different stimulus–response (S–R) ensembles and the different mapping assignments (from Fitts and Deininger, 1954; in the public domain).