

Human Category Learning: Implications for Backpropagation Models

Kruschke, J. K. (1993). Human category learning: implications for backpropagation models. *Connection Science*, 5, 3-36.

JOHN K. KRUSCHKE

(Received for publication 10 June 1992; revised paper accepted 29 October 1992)

Backpropagation (Rumelhart et al., 1986a) was proposed as a general learning algorithm for multi-layer perceptrons. This article demonstrates that a standard version of backprop fails to attend selectively to input dimensions in the same way as humans, suffers catastrophic forgetting of previously learned associations when novel exemplars are trained, and can be overly sensitive to linear category boundaries. Another connectionist model, ALCOVE (Kruschke 1990, 1992), does not suffer those failures. Previous researchers identified these problems; the present article reports quantitative fits of the models to new human learning data. ALCOVE can be functionally approximated by a network that uses linear-sigmoid hidden nodes, like standard backprop. It is argued that models of human category learning should incorporate quasi-local representations and dimensional attention learning, as well as error-driven learning, to address simultaneously all three phenomena.

KEYWORDS: Backpropagation, catastrophic forgetting, categorization, coarse coding, condensation, dimensional attention, error-driven learning, filtration, linear boundaries, local representation.

1. Introduction

Standard backpropagation (Rumelhart *et al.*, 1986a), or 'backprop', was originally proposed as a learning mechanism for multi-layer perceptrons (Rosenblatt, 1958; Minsky & Papert, 1969). Its main goal was to learn internal representations that could mediate complex mappings between inputs and outputs, as evidenced by the very title of the landmark report of Rumelhart *et al.* (1986a): 'Learning internal representations . . .'. Many papers have been written devoted to the analysis of the internal representations discovered by backprop (e.g. Elman, 1989; Hanson & Burr, 1991; Rosenberg, 1987).

Rumelhart *et al.* (1986a) did not address the question of whether backprop could model the course of human learning. This question can be asked at the neural or molar levels. It is generally (though not universally) agreed that backprop cannot be *directly* implemented in real neurons, given present-day knowledge of neural function (e.g. Grossberg, 1987; Rumelhart *et al.*, 1986b, p. 536; Stork,

J. K. Kruschke, Department of Psychology and Cognitive Science Program, Indiana University, Bloomington, IN 47405-4201, USA. E-mail: kruschke@ucs.indiana.edu.

1989). The general sentiment of most users of backprop was clearly expressed by Lehky and Sejnowski (1988, p. 454): "No biological significance is claimed for the algorithm (back propagation) by which the network developed, but, rather, the focus of interest is on the resulting mature network." Ultimately, neural plausibility is desirable for any model of behavior, especially for network models that are ostensibly 'brain style' and 'neurally inspired' (Rumelhart & McClelland, 1986). Nevertheless, there is a long history of learning models that make no attempt to contact neural functioning, although there is an implicit recognition that neural mechanisms must somehow implement them (e.g. Bower & Hilgard, 1981). The goal of such models is to capture accurately some of the molar learning behavior observed in people. Therefore, despite the neural implausibility of backprop, we can ask whether it reflects the course of learning at the molar, behavioral level.

Since 1986 many researchers have used backprop in models of human learning at the molar level. Several reports have emphasized its success (e.g. Cohen *et al.*, 1990; McClelland & Jenkins, 1991; Seidenberg & McClelland, 1989; Sejnowski & Rosenberg, 1988; Taraban *et al.*, 1989), and others have emphasized its failures (e.g. Gluck, 1991; McCloskey & Cohen, 1989; Pavel *et al.*, 1989; Ratcliff, 1990). This article isolates three failures of standard backprop to model human category learning. By 'category learning' I mean situations in which people learn to associate category labels with stimuli. First, backprop fails to learn category distinctions for which only a few stimulus dimensions are relevant faster than distinctions for which a large number of stimulus dimensions are relevant. In other words, backprop fails to attend selectively to stimulus dimensions the same way people do. Second, backprop suffers 'catastrophic forgetting' of previously learned associations when new associations are trained. Third, standard backprop learns linearly separable categorizations faster than non-linearly separable ones in some situations where people do not.

An alternative model, called ALCOVE (Kruschke, 1990, 1992), overcomes these problems. The model was motivated by a molar-level psychological theory, Nosofsky's (1986) generalized context model (GCM), rather than by neuron-like perceptrons. ALCOVE is closely related to the structure of standard backprop, in that it is also a feed-forward network that learns using gradient descent on error, but unlike backprop it has explicit attention strengths on the input dimensions, and it uses hidden nodes with a different activation function than used in backprop.

I would like to emphasize from the outset that the cause of backprop's problems is not the error-driven learning mechanism, but its particular architecture. One goal of this article is to demonstrate that these two models, though similar, generate very different behavior. I also show that backprop can be modified to mimic the functionality of ALCOVE, and then no longer suffers the three problems. Other researchers previously identified, qualitatively, the three problems focused upon in this article. What is new in this article is (i) the illustration of those problems with quantitative fits to robust, new data from simple (almost minimalist) human learning experiments, and (ii) an emphasis that the three problems exist simultaneously in standard backprop, so that solving one does not by itself make standard backprop a viable model of human category learning. Thus, the main goal of this article is to provide new emphasis and illustrations of these issues with quantitative fits to data from straightforward category learning experiments.

This article is organized as follows. I first describe the two models and point

out some general behavioral properties that can be gleaned from their structures. In the subsequent three sections, the models are applied to three situations in human category learning, demonstrating the three failures of standard backprop already mentioned. A modified version of backprop is then presented, which approximates ALCOVE and avoids the problems of standard backprop. The final section mentions other problems confronted by these models.

2. The Models

Both ALCOVE and standard backpropagation are feed-forward connectionist networks that learn by gradient descent on error. Thus, they both consist of a set of input nodes that encode the stimulus to be categorized, a set of output nodes that encode the category label, and a set of intermediate, 'hidden', nodes that transform the input representation into some internal representation. The layers of nodes are connected by weighted links, through which activation spreads from the input nodes, to the hidden nodes, to the category nodes. The models differ in two ways: (1) their particular choice of internal representation; and (2) whether or not there is a mechanism for dimensional attention learning.

2.1. ALCOVE

The architecture of ALCOVE was motivated by a molar-level psychological theory, Nosofsky's (1986) generalized context model (GCM). Like the GCM, ALCOVE assumes that input patterns can be represented as points in a multi-dimensional psychological space, as determined by multi-dimensional scaling (MDS) algorithms (Kruskal, 1964; Shepard, 1962). Thus, the first step in applying the model is determining the psychological coordinates of the stimuli. To do this, one obtains similarity ratings (or confusabilities) of pairs of stimuli, and determines the coordinate values in psychological space that best predict those similarities. This process is analogous to generating a spatial map of cities when all you are told is the distances (dissimilarities) between cities.

Each input node of ALCOVE encodes a single psychological dimension, with the activation of the node indicating the value of the stimulus on that dimension. Thus, if ψ_i is the psychological scale value of the stimulus on dimension i , then the activation of the i th input node is

$$a_i^{\text{in}} = \psi_i \tag{1}$$

Figure 1 shows the architecture of ALCOVE, illustrating a case with just two input dimensions.

Using MDS coordinates to encode the input does not introduce clandestine degrees of freedom into the model. Rather, the MDS solution is determined completely independently from the similarity ratings. Thus, the MDS representation *constrains* the model, and unlike some applications of standard backprop, we are not allowed to assume just any input representation that happens to be convenient. Of course, some other input representation might in fact work better, but the point remains that the MDS representation does not provide any extra degrees of freedom.¹

The i th input node is gated by a dimensional *attention strength* α_i . The attention strength on a dimension changes to reflect the relevance of that dimension for the

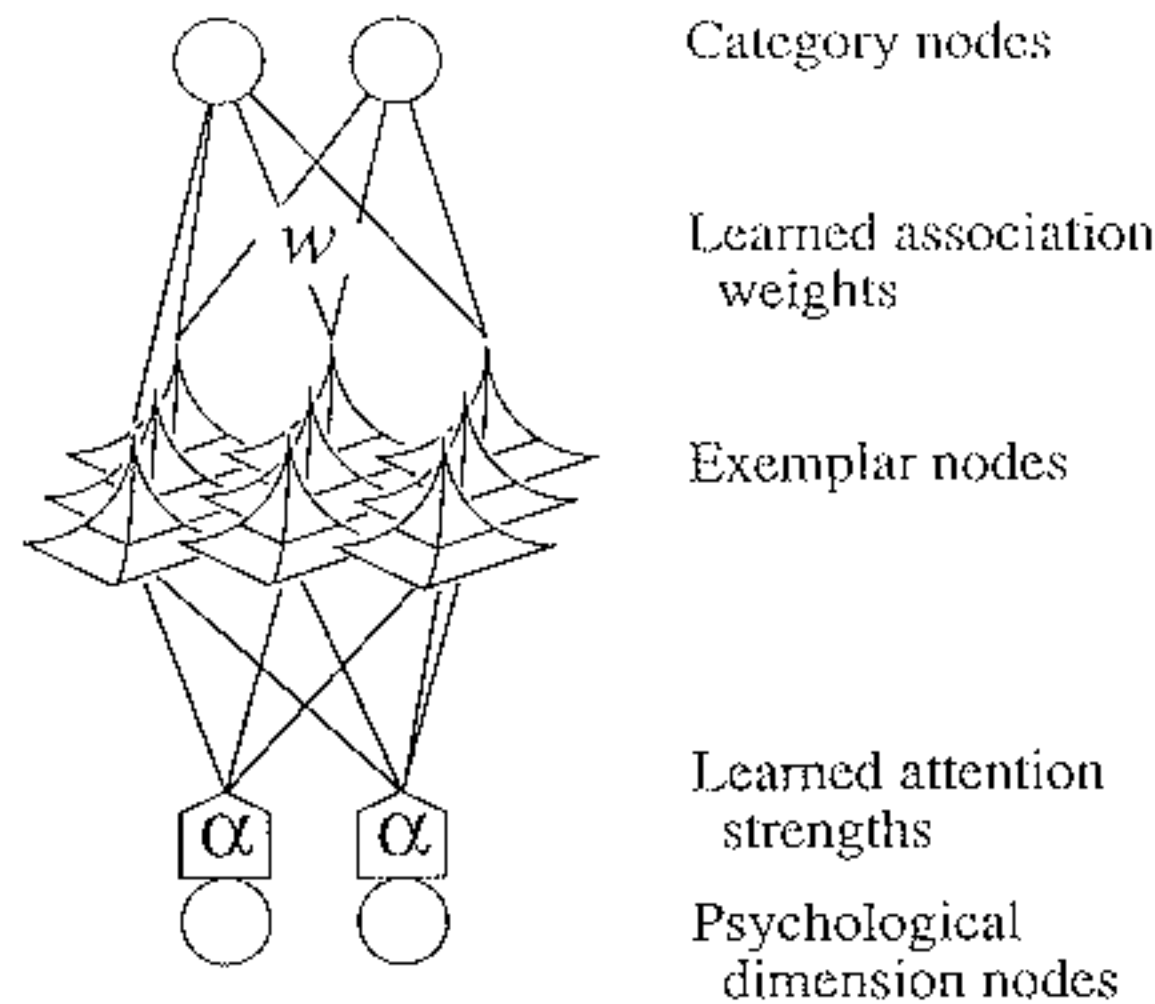


Figure 1. The structure of ALCOVE. The pyramids in the hidden layer indicate the activation profile of hidden nodes, as determined by equation (2), with $r = q = 1$.

particular categorization task at hand, as determined by a learning mechanism described below.

Each hidden node corresponds to a position in the multi-dimensional stimulus space, with one hidden node placed at the position of every training exemplar.² Each hidden node is activated according to the psychological similarity of the stimulus to the exemplar represented by the hidden node. The similarity function comes from the GCM and the work of Shepard (1962, 1987): Let the position of the j th hidden node be denoted (h_{j1}, h_{j2}, \dots) , and let the activation of the j th hidden node be denoted a_j^{hid} . Then

$$a_j^{\text{hid}} = \exp\left(-c\left(\sum_i \alpha_i |h_{ji} - a_i^{\text{in}}|^r\right)^{q/r}\right) \quad (2)$$

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 similarity metric and similarity gradient, respectively. For separable psychological dimensions, the city-block metric ($r = 1$) is used, while integral dimensions might call for a Euclidean metric ($r = 2$). An exponential similarity gradient ($q = 1$) is used here (Shepard, 1987).

When $r = 2$ in equation (2), the hidden nodes are a type of *radial basis function* (RBF), and ALCOVE can be construed as a type of radial basis function interpolation network (Broomhead & Lowe, 1988; Moody & Darken, 1989; Poggio & Girosi, 1990; Robinson *et al.*, 1988). Indeed, ALCOVE was born with the simple observation that the GCM can be implemented as an RBF network (Kruschke, 1990). Interestingly, a very similar model was independently invented by Hurwitz (1990), who was (quite differently) motivated by Estes's (1988) suggestion to combine exemplar representations with error-driven learning.

The dimensional attention strengths adjust themselves so that exemplars from different categories become less similar, and exemplars within categories become more similar. Consider a simple case of eight stimuli that form the corners of an octagon in a two-dimensional stimulus space, as shown in Figure 2. The stimuli are assigned to one of two categories, indicated by filled or open circles. Figure 2(a) shows a case in which one dimension can be ignored without loss of

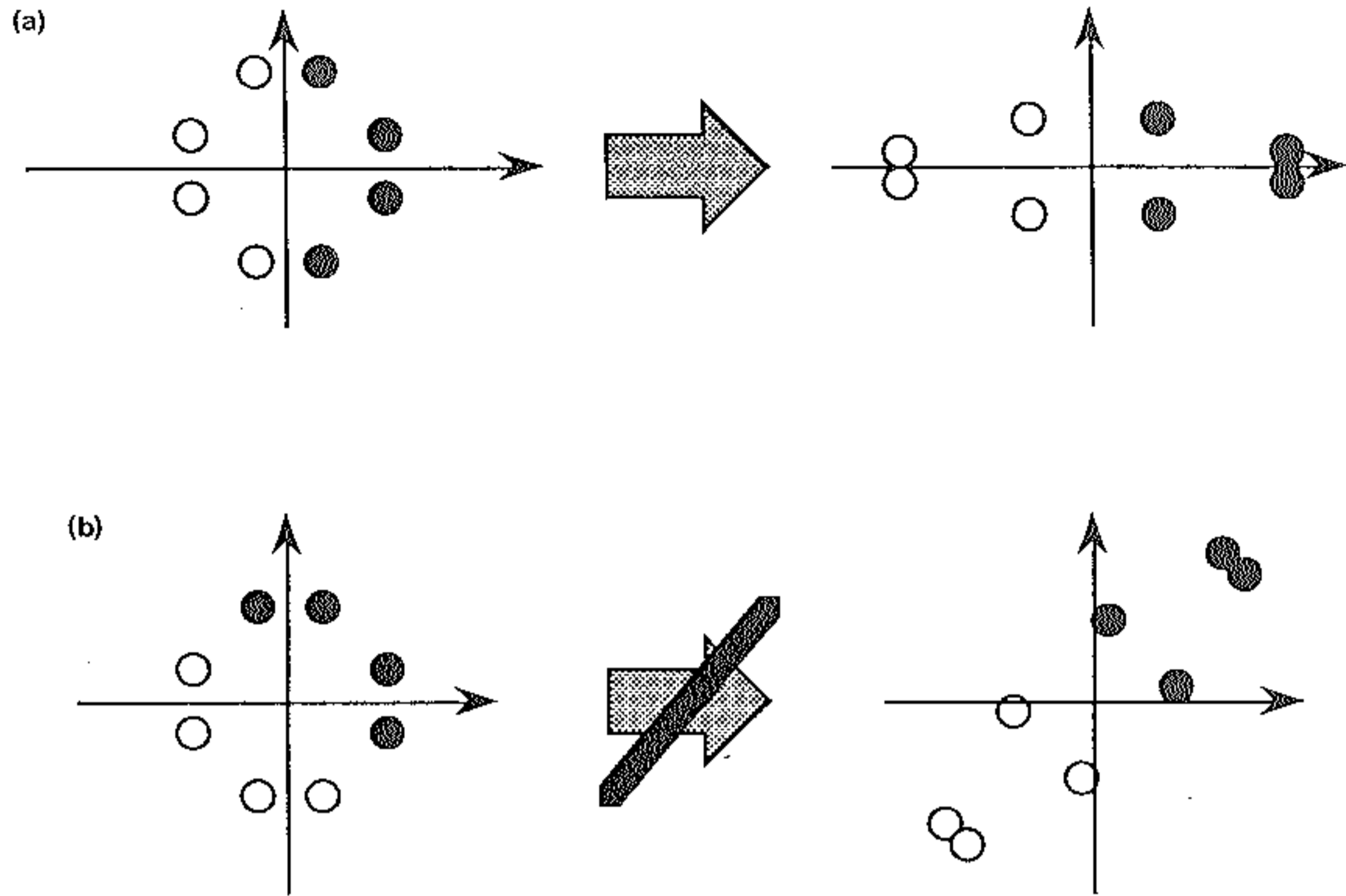


Figure 2. (a) Increasing attention to the horizontal dimension and decreasing attention to the vertical dimension causes exemplars of the two categories (denoted by filled and open circles) to have greater between-category dissimilarity and greater within-category similarity. (b) ALCOVE cannot differentially attend to diagonal axes.

classification accuracy. In this case, ALCOVE learns to increase the attention strength on the relevant dimension, and to decrease the attention strength on the irrelevant dimension. Increasing attention has the effect of stretching the dimension, and decreasing attention shrinks the dimension. Figure 2(b) shows a case in which neither dimension can be ignored without loss of classification accuracy. ALCOVE cannot stretch or shrink the stimulus space diagonally. As we will see, standard backprop does not share this anisotropy, and can differentially emphasize any direction in stimulus space. This difference has dramatic consequences for the models' predictions of the relative ease of learning such category structures, as will be demonstrated later.

Each hidden node in ALCOVE is connected to output nodes that correspond to response categories. The connection from the j th hidden node to the k th category node has a connection weight denoted w_{kj} , called the *association weight* between the exemplar and the category. The category nodes are activated by the linear rule used in the GCM and in the network models of Gluck and Bower (1988):

$$a_k^{\text{out}} = \sum_j^{\text{hid}} w_{kj} a_j^{\text{hid}} \quad (3)$$

Category activations are mapped to response probabilities using the same choice rule (Luce, 1963) as was used in the GCM and network models:

$$\Pr(K) = \exp(\varphi a_k^{\text{out}}) / \sum_k^{\text{out}} \exp(\varphi a_k^{\text{out}}) \quad (4)$$

where φ is a scaling constant. In other words, the probability of classifying the given stimulus into category K is determined by the magnitude of category K 's activation relative to the sum of all category activations.

Suppose the model is applied to the situation illustrated in Figure 2. In this case, there are two psychological dimensions, hence two input nodes; eight training exemplars, hence eight 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 (equation (1)). 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 (2). 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 (3). Finally, response probabilities are computed using equation (4).

The dimensional attention strengths, α_i , and the association weights, w_{kj} , are learned by gradient descent on sum-squared error, as used in standard backprop (Rumelhart *et al.*, 1986a) and in the network models of Gluck and Bower (1988). 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

$$E = \frac{1}{2} \sum_k^{\text{out}} (t_k - a_k^{\text{out}})^2 \quad (5)$$

where the teacher values are defined as

$$t_k = \begin{cases} \max(+1, a_k^{\text{out}}) & \text{if stimulus} \in k \\ \min(-1, a_k^{\text{out}}) & \text{if stimulus} \notin k \end{cases} \quad (6)$$

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 k th category, then the k th 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 an 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 to produce ultimately the same activations. Humble teachers are discussed further by Kruschke (1990, 1992), and they do not play a central role in this article.

Upon presentation of a training exemplar to ALCOVE, the association weights and attention strengths are changed by a small amount so that the error decreases. Following Rumelhart *et al.*, they are adjusted proportionally to the (negative of the) error gradient, which leads to the following learning rules, for $r = q = 1$ (derived in Kruschke, 1990, 1992):

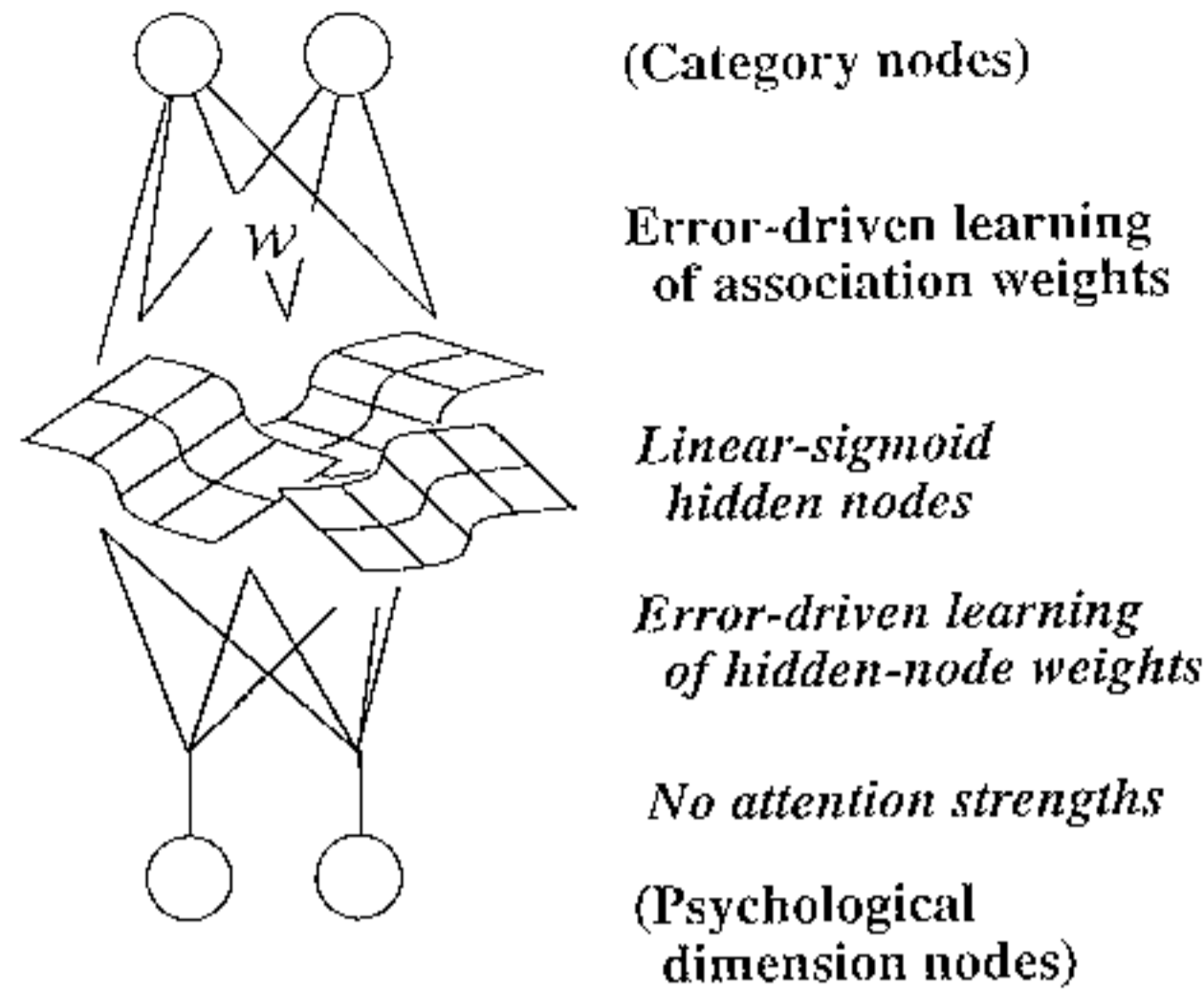


Figure 3. The structure of standard backpropagation.

$$\Delta w_{kj}^{\text{out}} = \lambda_w (t_k - a_k^{\text{out}}) a_j^{\text{hid}} \quad (7)$$

$$\Delta \alpha_i = -\lambda_\alpha \sum_{\text{hid } j} \left(\sum_{\text{out } k} (t_k - a_k^{\text{out}}) w_{kj} \right) a_j^{\text{hid}} c |h_{ji} - a_i^{\text{in}}| \quad (8)$$

where the λ s 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 attention strengths. If application of equation (8) gives an attention strength a negative value, then that strength is set to zero, because negative values have no psychologically meaningful interpretation.

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 weights and attention strengths are then adjusted according to equations (7) and (8).

In fitting ALCOVE to human learning data, there are four free parameters: the fixed specificity c in equation (2); the probability mapping constant ϕ in equation (4); the association weight learning rate λ_w in equation (7); and the attention strength learning rate λ_α in equation (8).

2.2. Standard Backpropagation

Standard backprop (Figure 3) uses *linear-sigmoid* nodes in its hidden layer, which have activation determined by

$$a_j^{\text{hid}} = 1 / \left(1 + \exp \left[-g \left(\sum_{\text{in } i} w_{ji}^{\text{hid}} a_i^{\text{in}} - \theta_j \right) \right] \right) \quad (9)$$

where g is a constant called the *gain* of the node (Kruschke & Movellan, 1991) and θ_j is the *threshold* of the node. The *linear-sigmoid* function was motivated as a generalized, or smoothed, version of the *linear-threshold* function in ‘neuron-like’ perceptrons.

The activation profiles of hidden nodes in ALCOVE and in backprop, as determined by equations (2) and (9), are shown in Figure 4. Note that the level

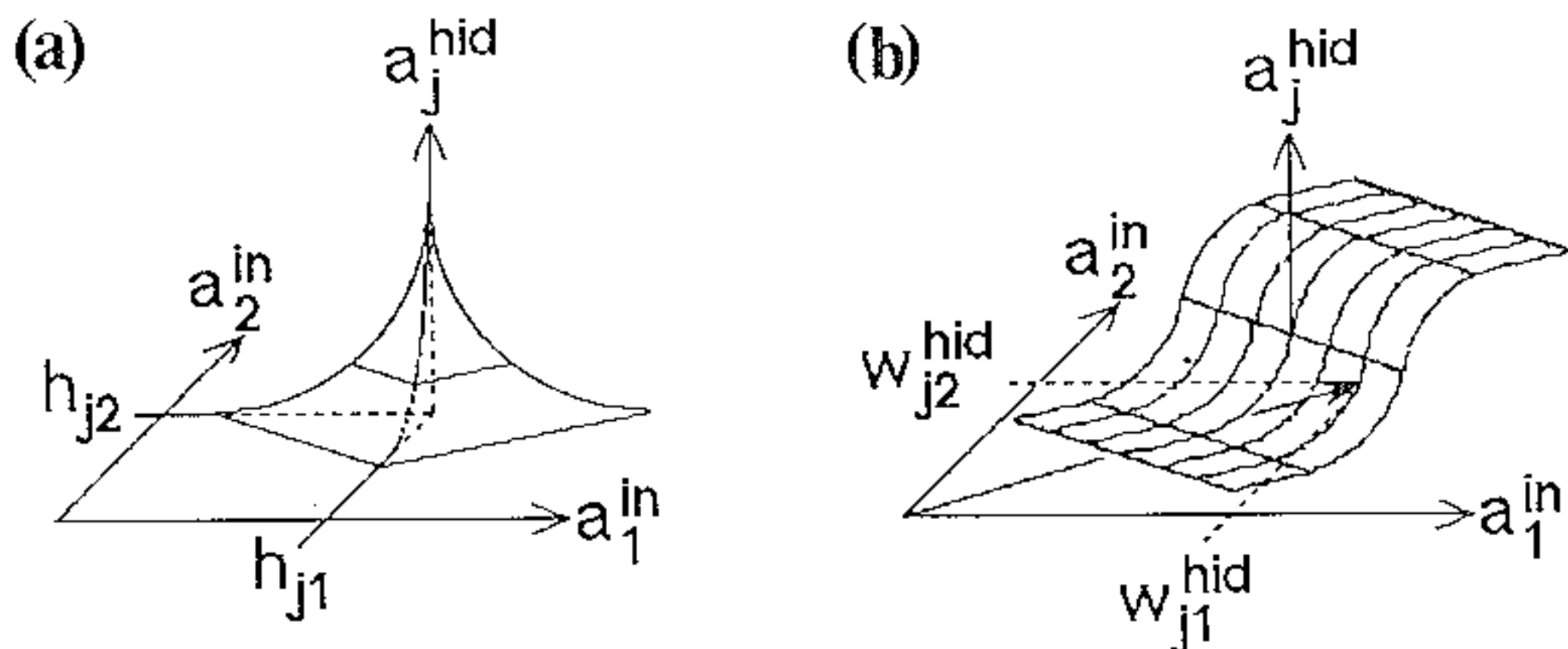


Figure 4. (a) Activation profile of a hidden node in ALCOVE (equation (2), with $r = q = 1$). (b) Activation profile of a hidden node in standard backpropagation (equation (9)). Hidden nodes in ALCOVE have a localized receptive field, whereas hidden nodes in backprop respond to an entire half-space.

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 (hyperplanes). Examples of level contours are shown in Figure 4 by the lines that mark 'horizontal' cross-sections through the activation profiles. In backprop, the *weights* in equation (9) determine the *orientation* of the linear level-contour in input space; the *threshold* in equation (9) determines the *distance* of the $a_j^{\text{hid}} = 0.50$ level contour from the origin; and the *gain* in equation (9) determines the *steepness* of the sigmoidal hill.

One important difference between the structures of ALCOVE and backprop is that the linear level-contours of the backprop node can be oriented in any direction in input space (depending on the hidden weights), whereas attention learning in ALCOVE can only stretch or shrink along the given input dimensions (recall the discussion accompanying Figure 2). In backprop, the linear level-contours can be equally easily aligned along the vertical or diagonal category boundaries in Figure 2. Consequently, backprop shows virtually no difference in learning speed between the two categorizations, as will be described at greater length below.

The gain parameter is not usually included in backpropagation (i.e. it is usually fixed at $g = 1.0$), but is included here for two reasons. First, it plays a role comparable to the specificity parameter in ALCOVE. Just as specificity determines the steepness of the generalization gradient for hidden nodes in ALCOVE, gain determines the steepness of the sigmoidal hill for hidden nodes in backprop. Second, it determines the effective range for initializing the random weights and thresholds of the hidden nodes. That is, the initial values of the weights and thresholds in equation (9) are drawn randomly from a uniform distribution on the interval $[-1, +1]$, and the gain acts as a multiplier to make the effective range $[-g, +g]$ (as can be seen by distributing g over the w_{ji}^{hid} and θ_j terms in equation (9)). This is important because it has been shown that the magnitude of initial weights can affect the behavior of backprop (Kolen & Pollack, 1990).

In order to make backprop and ALCOVE comparable in their output assumptions, the backprop network is given linear output nodes (equation (3)) with response probabilities determined by equation (4). The output weights are initialized to zero, for the same reason as in ALCOVE, viz, that initially there should be

no particular correspondence of stimuli (or their re-representation in the hidden layer) with category labels. The input layer of backprop is also assumed to use the same representation used by ALCOVE, because backprop makes no particular assumptions about the input presentation. The upshot is that there are two critical distinctions between ALCOVE and backprop: the difference in hidden node activation functions; and the presence or absence of a dimensional attention learning mechanism.

Learning in backprop proceeds as follows. Upon presentation of a training exemplar, the weights and thresholds are adjusted proportionally to the (negative of the) error gradient, which leads to the following learning rules (for derivations see Kruschke & Movellan, 1991; Rumelhart *et al.*, 1986a):

$$\Delta w_{kj}^{\text{out}} = \lambda_{\text{out}}(t_k - a_k^{\text{out}})a_j^{\text{hid}} \quad (10)$$

$$\Delta w_{ji}^{\text{hid}} = \lambda_{\text{hid}} \left(\sum_{k \text{ out}} (t_k - a_k^{\text{out}}) w_{kj}^{\text{out}} \right) (1 - a_j^{\text{hid}}) a_j^{\text{hid}} g a_i^{\text{in}} \quad (11)$$

$$\Delta \theta_j = -\lambda_{\theta} \left(\sum_{k \text{ out}} (t_k - a_k^{\text{out}}) w_{kj}^{\text{out}} \right) (1 - a_j^{\text{hid}}) a_j^{\text{hid}} g \quad (12)$$

The same learning rate, λ_{out} , applies to all the output weights. Likewise, there is only one learning rate, λ_{hid} , for all the hidden weights, and one learning rate, λ_{θ} , for all the hidden thresholds. Note that equation (10) is the same as equation (7), because the output layers of the two models are assumed to have the same structure. Note also the similarity of equations (11) and (12) to equation (8).

In fitting backprop to data there are five parameters: the three learning rates in equations (10), (11) and (12); the gain g , and the choice probability constant ϕ .

2.3. Summary of Models

Throughout this article, when I use the term ‘backprop’ I am referring to the use of linear-sigmoid hidden nodes as in standard backpropagation. When I mean to refer to the learning mechanism, I will call it gradient descent on error, or error-driven learning. The target of this article is linear-sigmoid hidden nodes and dimensional attention, not gradient descent on error.

ALCOVE can be construed as a coalescence of three intellectual currents. First, it uses an exemplar-based internal representation that stems directly from theories of similarity-based generalization and attentionally weighted stimulus dimensions proposed by Shepard (1957, 1987), Medin and Schaffer (1978), Estes (1986) and Nosofsky (1986). Second, it uses error-driven learning, as in the models of Rescorla and Wagner (1972), Rumelhart *et al.* (1986a/b), and Gluck and Bower (1988). Third, it is essentially a form of radial-basis function interpolation network, as described by Broomhead and Lowe (1988), Robinson *et al.* (1988), Moody and Darken (1989), Poggio and Girosi (1990), and others. Thus ALCOVE should not be construed as a model opposed to standard backprop, but rather as a variant of it.

3. Lack of Selective Attention

As described in the previous section, backprop and ALCOVE differ in two critical ways: the shapes of their hidden node receptive fields, and the presence or absence of learned dimensional attention strengths. This section will concentrate on the dimensional attention strengths, and subsequent sections will focus on the receptive fields.

3.1. *Filtration vs Condensation*

Imagine attending a display of fireworks, and hearing the crowd respond "ooh!" to some and "ahh!" to others. Let's suppose that the responses are not random, but are determined by some visible properties of each burst, and that it is now our task to *learn* which bursts get which response. Suppose that all red fireworks elicit an "ooh!", while all other colors conjure an "ahh!" It is intuitively plausible that one would quickly learn to attend to the dimension of color, and ignore other stimulus dimensions. Suppose instead that the cheer for a burst could only be determined by the combination of two (or more) dimensions, such as color and size. It seems likely that accurate classification learning would take longer. The first situation, in which categorization of a stimulus could be accomplished by attending to just a subset of the available stimulus dimensions, has been called a 'gating' or 'filtration' task, because the irrelevant dimensions can be filtered or gated away, without loss of classification accuracy. The second situation, in which more than one dimension must be attended for accurate classification, has been called a 'condensation' task, because information from more than one dimension must be condensed into a single classification decision (Garner, 1974; Posner, 1964).

It has been well established that filtration tasks are indeed easier than condensation tasks (e.g. Garner, 1974; Gottwald & Garner, 1972, 1975; Kemler & Smith, 1978; Posner, 1964), confirming the intuition in the hypothetical case of classifying fireworks. The advantage of filtration over condensation can be appraised as a robust and fundamental phenomenon that models of category learning should address.

Explanations of filtration advantage have invoked the notion of selective attention. For example, Garner (1974) argued that condensation was difficult because the subject could not attend to a combination of dimensions with the same efficiency as he or she could (selectively) attend to a single dimension. It seems only natural, then, that models of category learning should include mechanisms for selective attention to stimulus dimensions. In this section I show that backprop does not have an appropriate form of selective attention, and fails to show filtration advantage. On the other hand, ALCOVE has dimensional attention strengths built in, and does exhibit filtration advantage.

In order to compare directly the models' quantitative predictions, new data had to be obtained from filtration and condensation categorization tasks. Unfortunately, pyrotechnic displays are unwieldy stimuli, so instead I used category exemplars consisting of rectangles that varied in height, with an interior segment that varied in its lateral position (Figure 5). Different groups of subjects were trained on the four category structures shown in Figure 6. Only eight different stimuli were presented, four of which were assigned to one category and the remaining four to the other category, indicated in Figure 6 by blank and filled circles. There were two filtration conditions, one in which height was relevant and

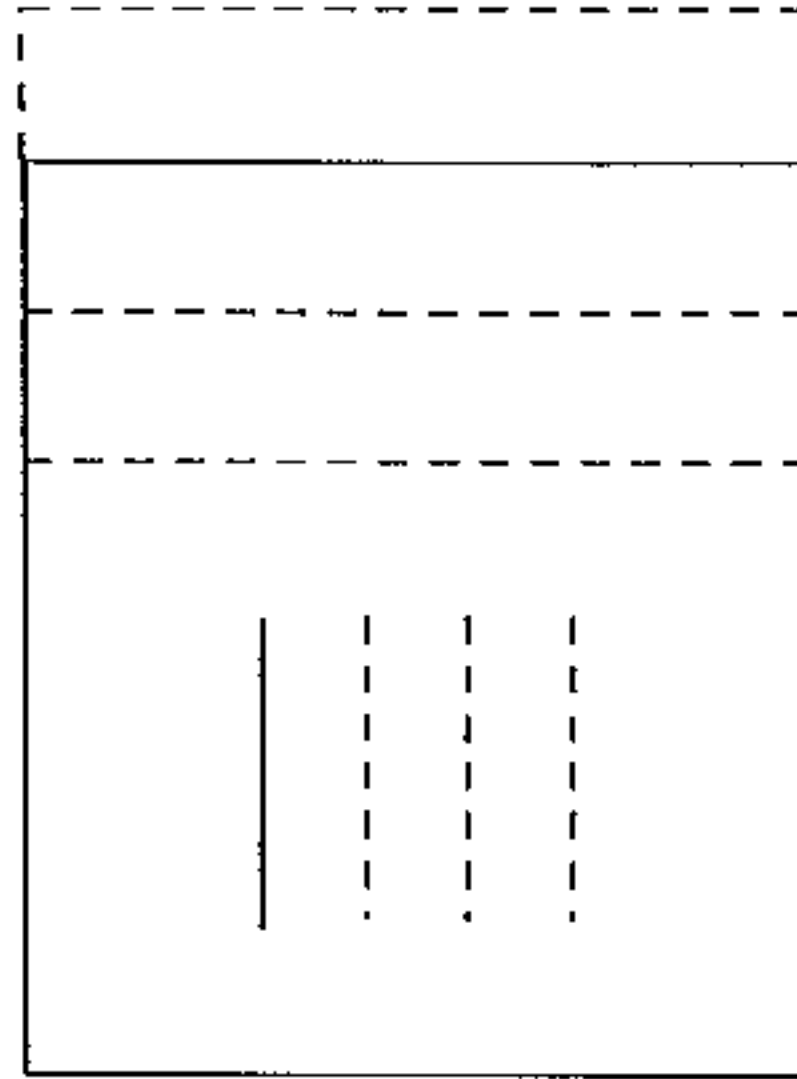


Figure 5. Stimuli. Solid lines show one combination of rectangle height and lateral position of interior segment. Dotted lines show alternative heights and positions.

one in which position was relevant (Figure 6(a)), and two condensation conditions (Figure 6(b)).

I chose the structures in Figure 6 primarily because they are (very nearly) rotations of each other, and standard backprop is insensitive to rotations of the input space. A second motive for these structures was that the clustering of exemplars, considered alone, predicts that the condensation situations should be at least as easy the filtration conditions. Exact quantitative predictions on the basis of clustering depend on how one wishes to define a measure of clustering. As an illustration, suppose we use the mean city-block separation of exemplars within categories vs between categories (Medin & Schwanenflugel, 1981), and let the

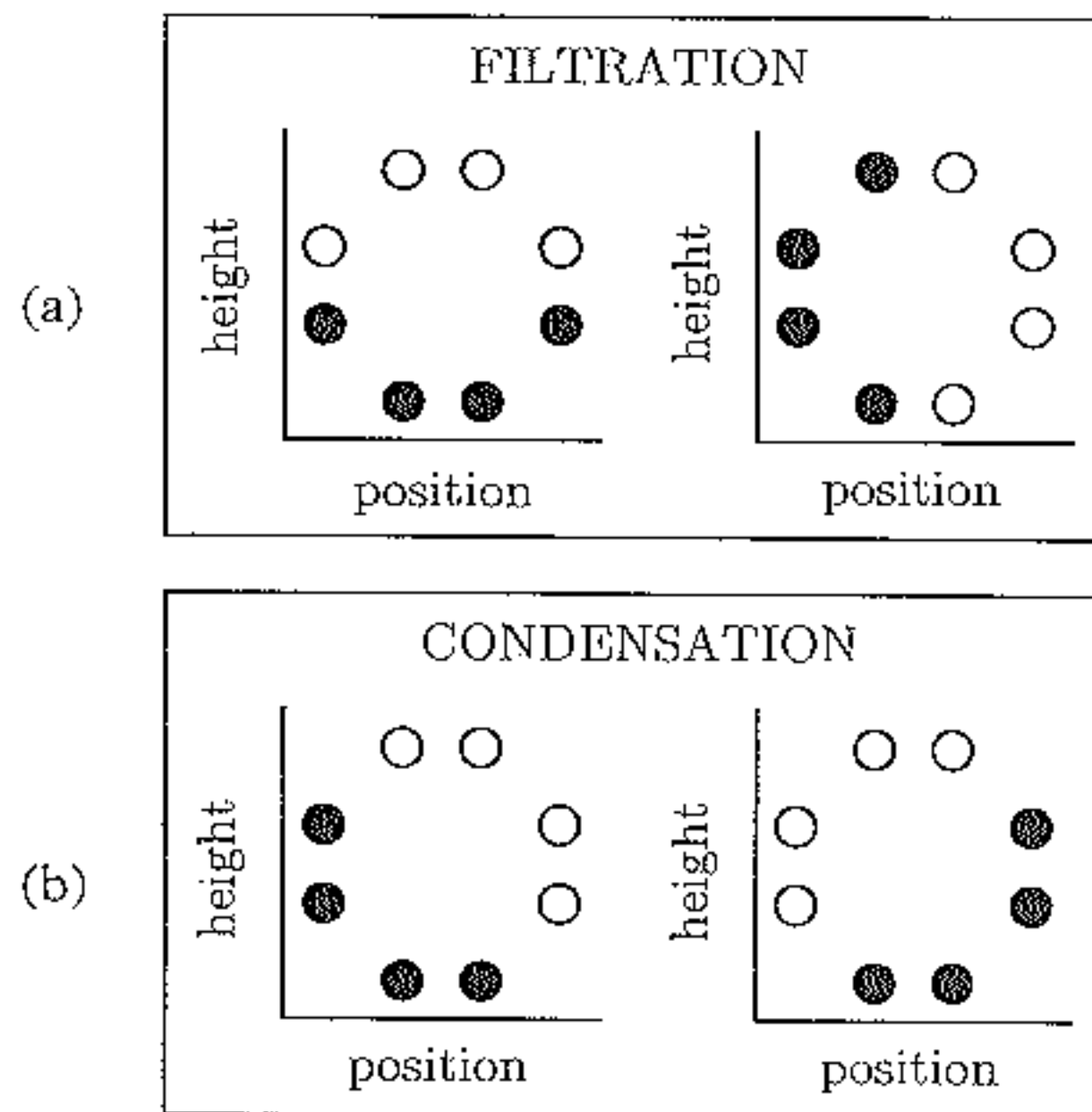


Figure 6. Structure of the filtration and condensation categories. Open circles denote one category, filled circles the other.

distance between levels on each dimension be 1 scale unit. Then both the filtration and condensation structures have a mean within-category separation of 2.33 and a mean between-category separation of 3.25. If city-block separations are (non-linearly) converted to similarities using the exponential function of distance in equation (2) (with $c = r = q = 1$), then the condensation structure should be easier than the filtration structure, as the mean similarity of exemplars within categories is 0.16 for condensation but only 0.13 for filtration, and the mean similarity of exemplars between categories is only 0.049 for condensation but 0.074 for filtration. A third consideration in the choice of these structures was that the same exemplars are used for both the filtration and condensation tasks; only the category assignment changes. Thus any differences between the filtration and condensation categories can be attributed to their structures rather than to effects of individual exemplars.

3.1.1. Procedure. Stimuli were presented with a PC-clone computer using VGA resolution, as white lines against a black background. Viewing distance was about 0.9 m, so that the height of the tallest rectangle subtended about 13 degrees of visual angle. The rectangles were presented so that the lower horizontal line was in the same position on every trial, centred horizontally on the screen. Of the 16 possible combinations of dimension values, only eight were used, corresponding to the abstract structure in Figure 2. In all experiments reported in this article, subjects were run individually in a dimly lit, quiet booth. All experiments were programmed using Micro Experimental Laboratory (Schneider, 1988).

Instructions were presented to the subject on the computer screen, and read aloud by the experimenter. Subjects were told that they must learn which stimulus belonged to which category. The category labels were 'B' and 'N'. Subjects responded by using the index and middle fingers of their dominant hand to press the corresponding keys on the keyboard. The instructions said that the stimuli varied on just two dimensions, height and position. As part of the instructions, subjects were shown all the stimuli without any category feedback (and without any responses from the subject). Each of the eight stimuli was shown twice, in a random sequence that was the same for all subjects. Subjects were instructed that there was no emphasis on response speed, and that they had up to half minute to respond on each trial.

Each training trial consisted of a presentation of a stimulus, which was terminated when the subject pressed a response key, followed by 1000 ms feedback indicating whether the response was 'correct' or 'wrong' accompanied by a 333 ms tone if wrong, followed by 750 ms feedback indicating the correct response ('That was a B' or 'That was an N').

The four category distinctions were given to different groups of subjects. Category labels were counterbalanced within groups, so that a given stimulus had correct label 'B' for half the subjects and 'N' for the other half. Every subject in every group saw the same fixed sequence of stimuli; all that varied between groups was the category labels assigned to the stimuli. There were 64 uninterrupted training trials. The experiment lasted about 20 minutes.

3.1.2. Subjects. A total of 160 subjects participated, 40 in each category type, for partial credit in an introductory psychology course.

