Categories and Concepts

Computational models (part 1)

Brenden Lake

PSYCH-GA 2207

Why build a computational model?

"Verbally expressed statements are sometimes flawed by internal inconsistencies, logical contradictions, theoretical weaknesses and gaps. A running computational model, on the other hand, can be considered as a sufficiency proof of the internal coherence and completeness of the ideas it is based upon."
(Fum, Del Missier, & Stocco, 2007)

Some famous psychological theories…

- Attention is like a spotlight
- A child learning about the world is like a scientist theorizing about science
- Language influences thought
- Working memory is having 7 +/- 2 slots to store information
- Categorization happens by comparing novel instances to past exemplars
- Categories influence perception

Each of these theories benefits from formalization with a computational model to…

- **Make predictions explicit**
- Implications often **defy expectations**
- **Aid communication** between scientists
- Support **cumulative progress**

My view echos Richard Feynman’s

“What I cannot create, I do not understand.”

Agenda for the next 2 classes

Two influential computational models

- ALCOVE
  - “An exemplar-based connectionist model of category learning”, Krushcke
- Rational Model
  - “An exemplar-based connectionist model of category learning”, Anderson

And some comments on category learning in modern AI

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.

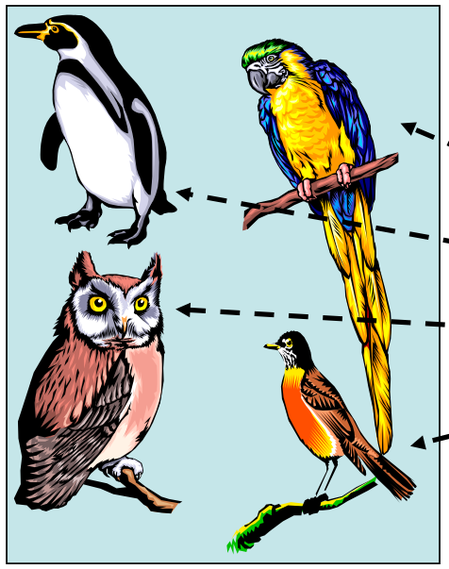
This article describes a connectionist model of category learning called ALCOVE (attention learning covering map). Any model of category learning must address the two issues of what representation underlies category knowledge and how that representation is used in learning. ALCOVE combines the exemplar-based 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 gener-

dient 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

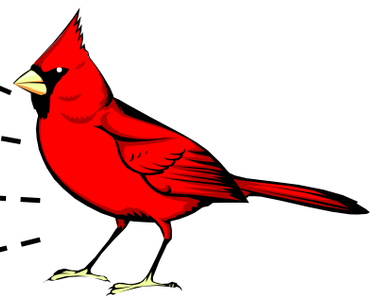
Like the Context Model, ALCOVE is an exemplar model

$x \in C$



Birds
You've Seen

y



Bird?

Review of the "Context Model"

Similarity between two items *x* and *y*
(Medin & Schaffer, 1978)

Category A **sim**(y, x) **Category B**

	D_1	D_2	D_3	D_4
x	1	1	1	1
	1	1	1	0
	0	0	0	1

0.09

D_1	D_2	D_3	D_4
0	0	0	0
0	0	1	1
1	1	0	0

Transfer item

y	0	1	0	1

$$\textbf{sim}(y, x) = \prod_{D_i} m^{1[x_i \neq y_i]} = m \cdot 1 \cdot m \cdot 1 = 0.09$$

"mismatch" free parameter such that $m = 0.3$

Review: Context model

"classification is based on similarity to all exemplars in a class"
(Medin & Schaffer, 1978)

Item similarity

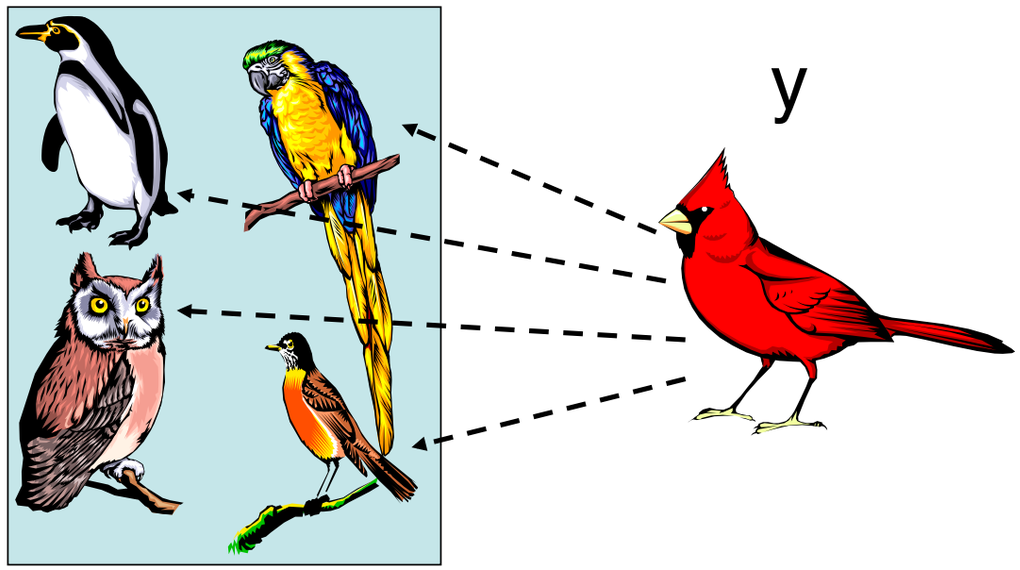
$$\textbf{sim}(y, x) = \prod_{D_i} m^{1[x_i \neq y_i]}$$

"mismatch" free parameter *m*

Category similarity

$$\textbf{sim}(y, C) = \sum_{x \in C} \textbf{sim}(y, x)$$

$x \in C$



Birds
You've Seen

Bird?

Probability of classification

$$P(y \in C) = \frac{\textbf{sim}(y, C)}{\sum_{C'} \textbf{sim}(y, C')}$$

ALCOVE's similarity between two items *x* and *y*

Context model's item similarity

$$\textbf{sim}(y, x) = \prod_{D_i} m_i^{1[x_i \neq y_i]}$$

(assuming we have a different mismatch parameter m_i for each dimension)

Recall this identity:

$$\prod_{i=1}^{N} z_i = \prod_{i=1}^{N} e^{\log z_i} = e^{\sum_{i=1}^{N} \log z_i}$$

$$\alpha_i = \log m_i$$

(equivalent for binary dims, with this identity)

ALCOVE's item similarity (**also works for continuous features**)

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \log m_i \; 1[x_i \neq y_i]} = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

(parameters *c*, *q*, and *r* not shown)

ALCOVE's similarity between two items *x* and *y*

Category A

D_1	D_2	D_3	D_4
1	1	1	1
1	1	1	0
0	0	0	1

x

sim(y, x)

0.09

Category B

D_1	D_2	D_3	D_4
0	0	0	0
0	0	1	1
1	1	0	0

Transfer item

y

0	1	0	1

$$\textbf{sim}(y, x) = \prod_{D_i} m^{1[x_i \neq y_i]} = m \cdot 1 \cdot m \cdot 1 = 0.09$$

"mismatch" free parameter such that $m = 0.3$

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \log m_i \, |x_i - y_i|} = e^{-1.20 + 0 \, -1.20 + 0} = 0.09$$

ALCOVE

Item similarity
(discrete or continuous)

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

key parameter: attention weights α_i

Category similarity

$$\textbf{sim}(y, C) = \sum_{x \in X} w_{cx}\ \textbf{sim}(y, x)$$

X is the set of all exemplars across all categories

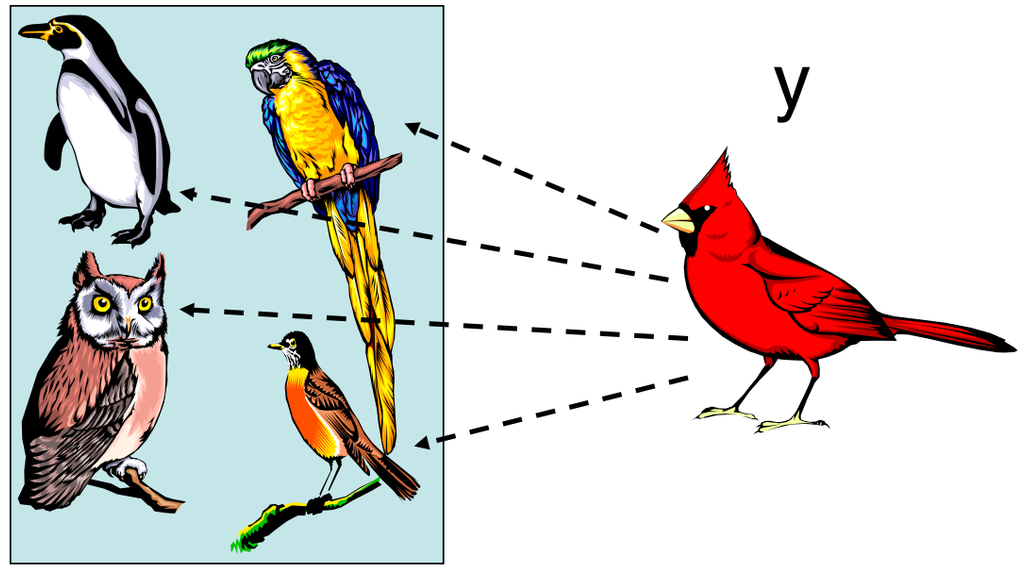
key parameter: association weights between exemplar *x* and category *c* w_{cx}

Probability of classification

$$P(y \in C) = \frac{e^{\phi \textbf{sim}(y,C)}}{\sum_{C'} e^{\phi \textbf{sim}(y,C')}}$$

ϕ response "temperature"

$x \in X$



Birds
You've Seen

Bird?

ALCOVE

Context model

Item similarity

$$\textbf{sim}(y,x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

key parameter: attention weights α_i

$$\textbf{sim}(y,x) = \prod_{D_i} m^{1[x_i \neq y_i]}$$

"mismatch" free parameter *m*

Category similarity

$$\textbf{sim}(y,C) = \sum_{x \in X} w_{cx}\, \textbf{sim}(y,x)$$

key parameter: association weights w_{cx} between exemplar *x* and category *c*

$$\textbf{sim}(y,C) = \sum_{x \in C} \textbf{sim}(y,x)$$

(only considers exemplars *x* assigned to C)

Probability of classification

$$P(y \in C) = \frac{e^{\phi \textbf{sim}(y,C)}}{\sum_{C'} e^{\phi \textbf{sim}(y,C')}}$$

ϕ response "temperature"

$$P(y \in C) = \frac{\textbf{sim}(y,C)}{\sum_{C'} \textbf{sim}(y,C')}$$

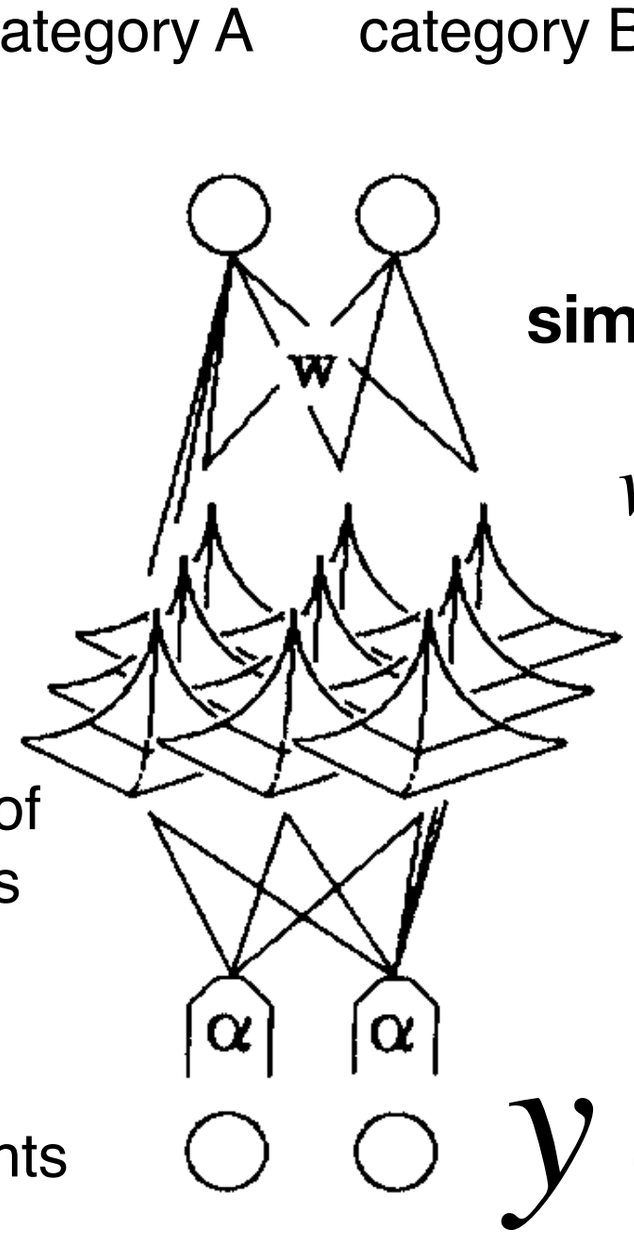
ALCOVE

Response rule

$$P(y \in A) = \frac{e^{\phi \mathbf{sim}(y,A)}}{e^{\phi \mathbf{sim}(y,A)} + e^{\phi \mathbf{sim}(y,B)}}$$

$$\mathbf{sim}(y, x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

x denotes one of the exemplars



$$\mathbf{sim}(y, C) = \sum_{x \in X} w_{cx}\ \mathbf{sim}(y, x)$$

w_{cx} association weights between exemplar *x* and category *c*

α_i attention weights

y current stimulus

How does ALCOVE learn?

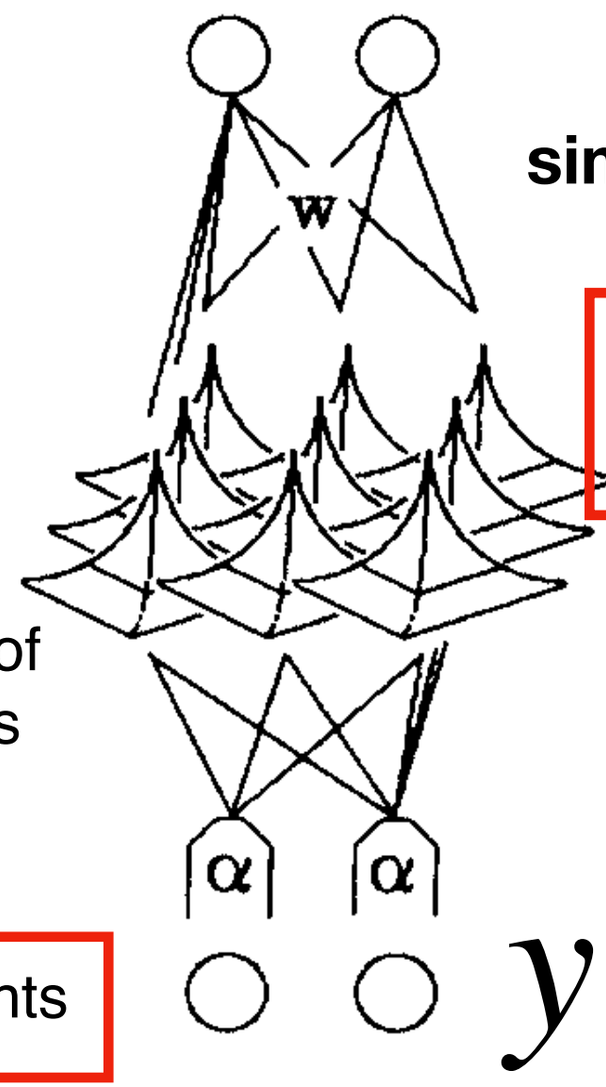
Learning is incremental fitting of the attention weights and association weights

Response rule

$$P(y \in A) = \frac{e^{\phi \textbf{sim}(y,A)}}{e^{\phi \textbf{sim}(y,A)} + e^{\phi \textbf{sim}(y,B)}}$$

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

x denotes one of the exemplars



$$\textbf{sim}(y, C) = \sum_{x \in X} w_{cx}\ \textbf{sim}(y, x)$$

w_{cx} association weights between exemplar *x* and category *c*

α_i attention weights

y current stimulus

Learning by optimizing an objective function

Let's use log-likelihood as our **objective function**:

$$F(w, \alpha) = \sum_{x \in A} \log P_{w,\alpha}(x \in A) + \sum_{x \in B} \log(P_{w,\alpha}(x \in B))$$

("maximum likelihood" is also the objective function for linear regression, logistic regression, and many other stats/machine learning algorithms…)

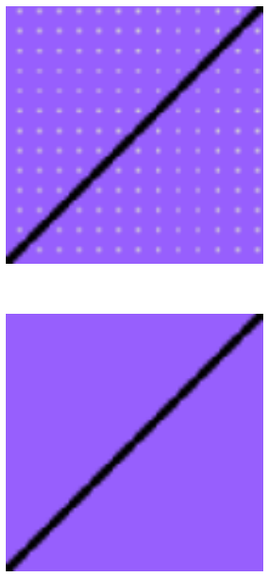
Then the optimal parameters are given by:

$$\textbf{argmax}_{w,\alpha}\ F(w, \alpha)$$

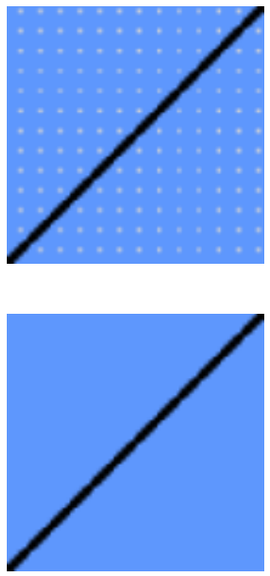
(note that this is a different objective function than used in original ALCOVE paper)

ALCOVE applied to SHJ Type I category

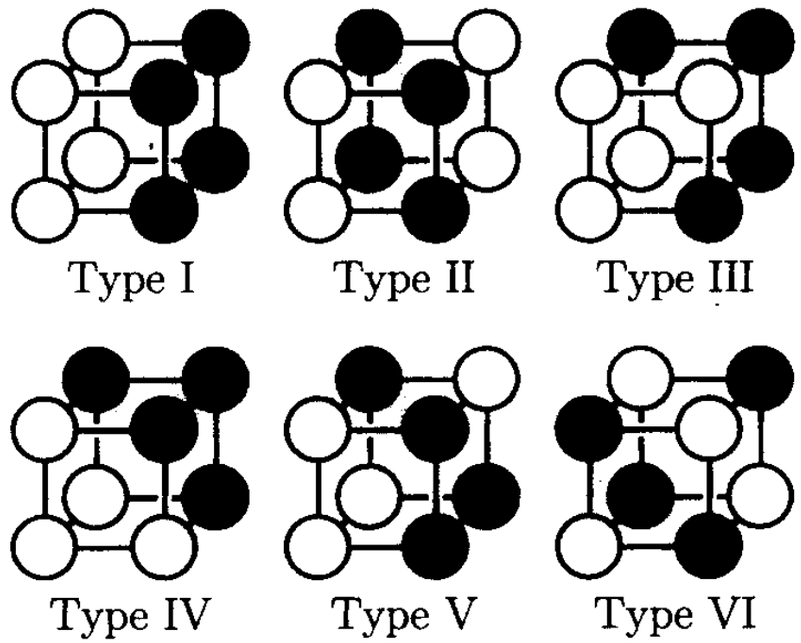
category A

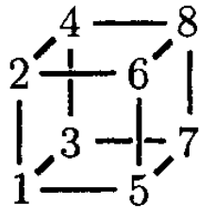


category B



Shepard, Hovland, & Jenkins (1961)





dim 3 **slash**

dim 2 **texture**

dim 1 **color**

Network BEFORE training

response $P(y \in A)$ 0.5

$$\textbf{sim}(y, C) = \sum_{x \in X} w_{cx}\ \textbf{sim}(y, x)$$

association weights w_{Ax} 0.0 0.0 0.0 0.0

$\textbf{sim}(y, x)$ 1.0 0.11 0.11 0.01

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

w_{Ax} 0.0 0.0 0.0 0.0

$\textbf{sim}(y, x)$ 0.11 0.01 0.01 0.0

} Exemplars X

attention weights α_{color} 0.33 $\alpha_{texture}$ 0.33 α_{slash} 0.33

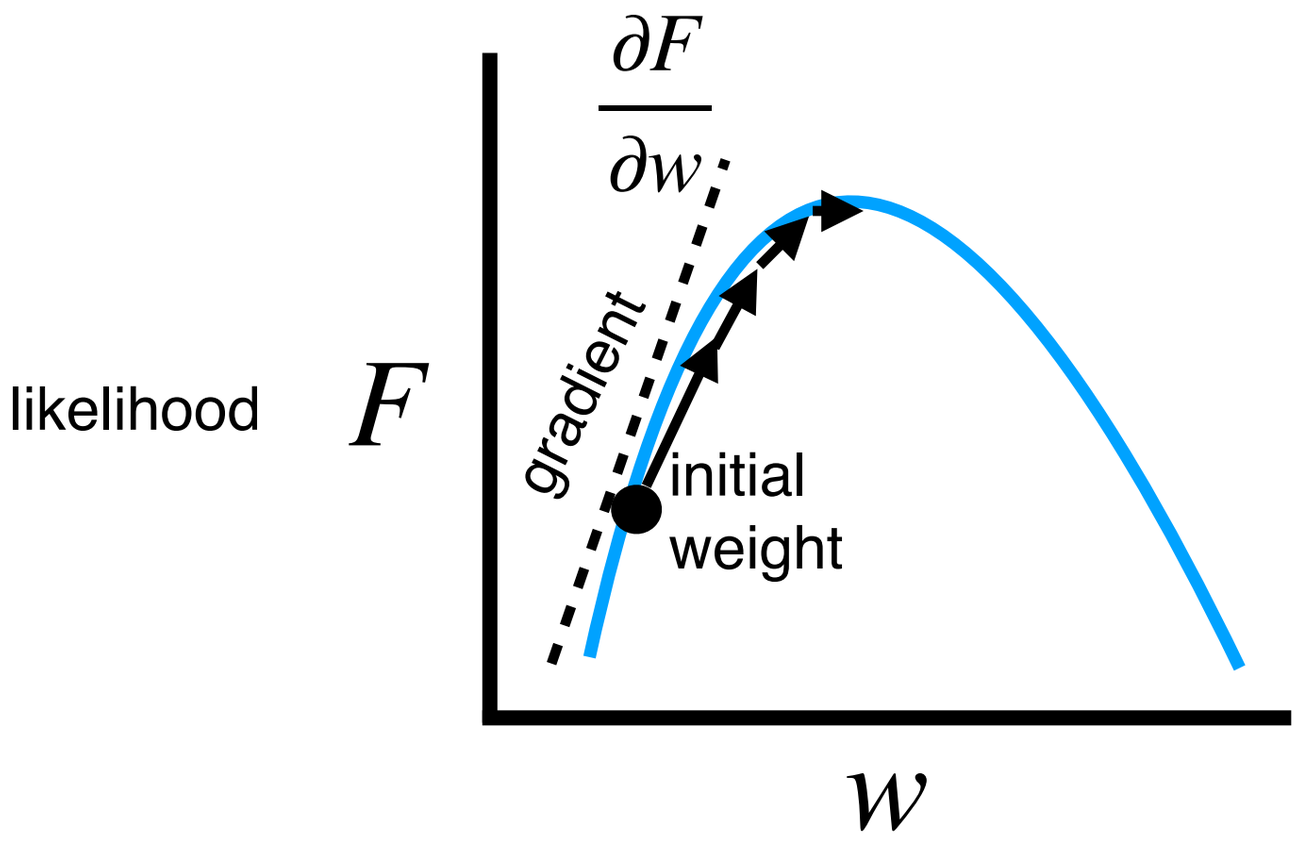
current stimulus y

Maximizing likelihood via gradient descent

(the workhorse of modern machine learning, and often cognitive modeling too)

objective function to maximize

$$F(w, \alpha) = \sum_{x \in A} \log P_{w,\alpha}(x \in A) + \sum_{x \in B} \log(P_{w,\alpha}(x \in B))$$



Computing the gradient tells us which direction to go for steepest ascent:

$$w \leftarrow w + \gamma \frac{\partial F}{\partial w}$$

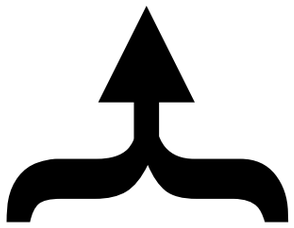
γ : learning rate (small constant)

ALCOVE after one gradient step

response $P(y \in A)$ **0.503**

Response is a little better, since y is in A!

$$\textbf{sim}(y, C) = \sum_{x \in X} w_{cx}\ \textbf{sim}(y, x)$$



association weights w_{Ax} 0.015 0.002 0.002 0.0

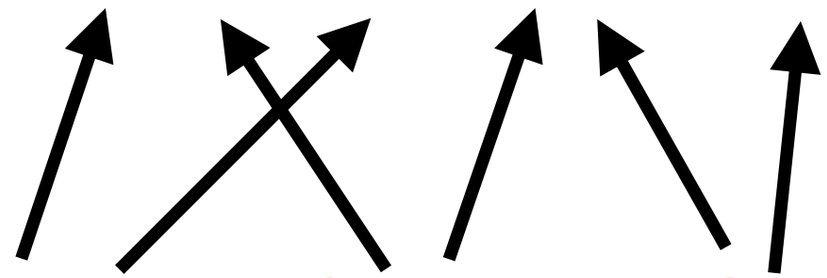
sim(y, x) **1.0** **0.11** **0.11** **0.01**

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

w_{Ax} 0.002 0.0 0.0 0.0

sim(y, x) **0.11** **0.01** **0.01** **0.0**

} Exemplars X



attention weights α_{color} 0.33 $\alpha_{texture}$ 0.33 α_{slash} 0.33

current stimulus y

ALCOVE after many more gradient steps…

response $P(y \in A)$ 0.99

Response is now nearly perfect

$$\textbf{sim}(y, C) = \sum_{x \in X} w_{cx}\ \textbf{sim}(y, x)$$

association weights w_{Ax} 1.15 1.14 1.14 1.14

sim(y, x) 1.0 1.0 1.0 1.0

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

w_{Ax} -1.14 -1.14 -1.14 -1.14

sim(y, x) 0.03 0.03 0.03 0.03

} Exemplars X

attention weights

α_{color}	$\alpha_{texture}$	α_{slash}
0.544	0.0	0.0

Attention only looks at color

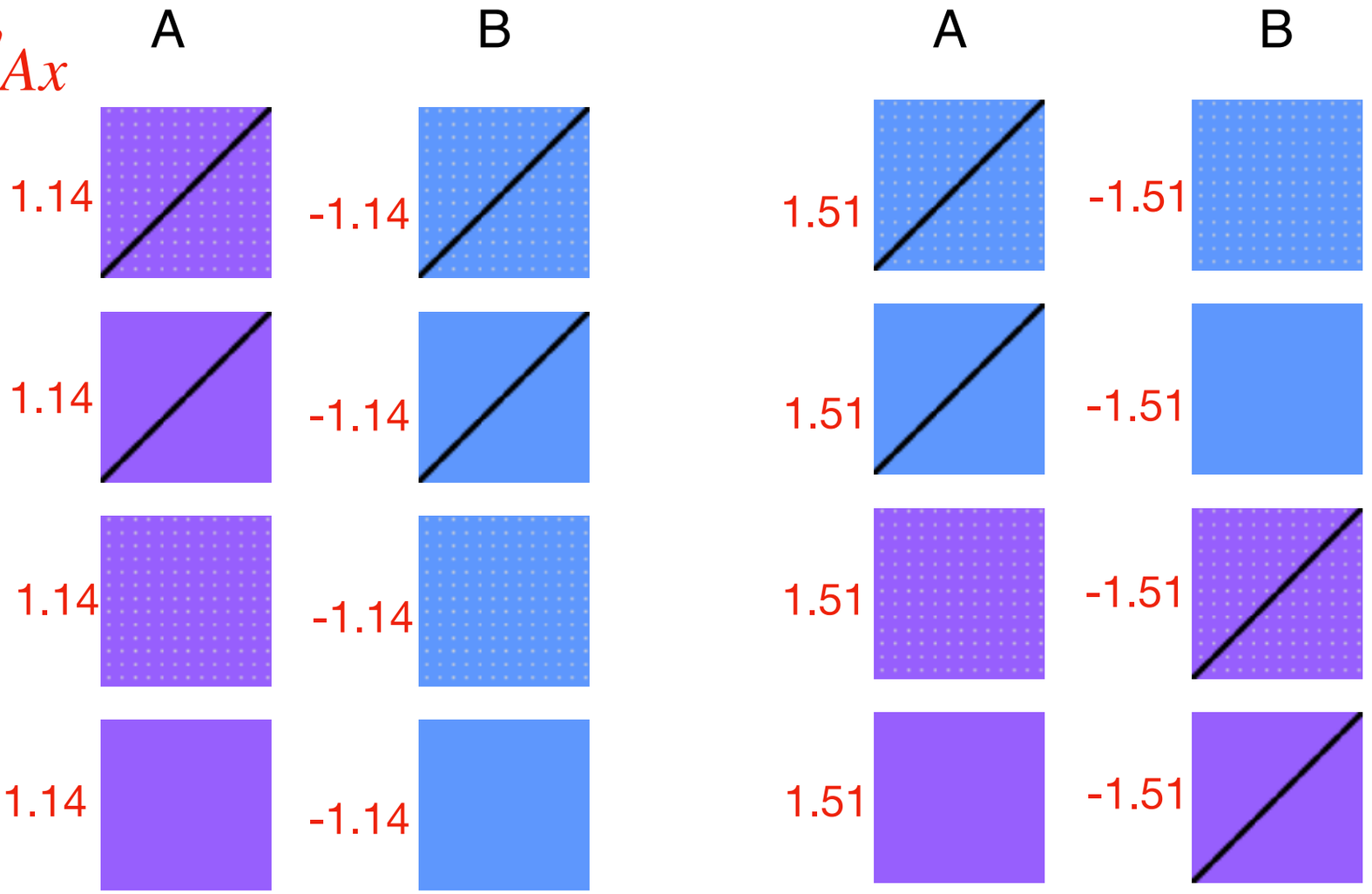
current stimulus y

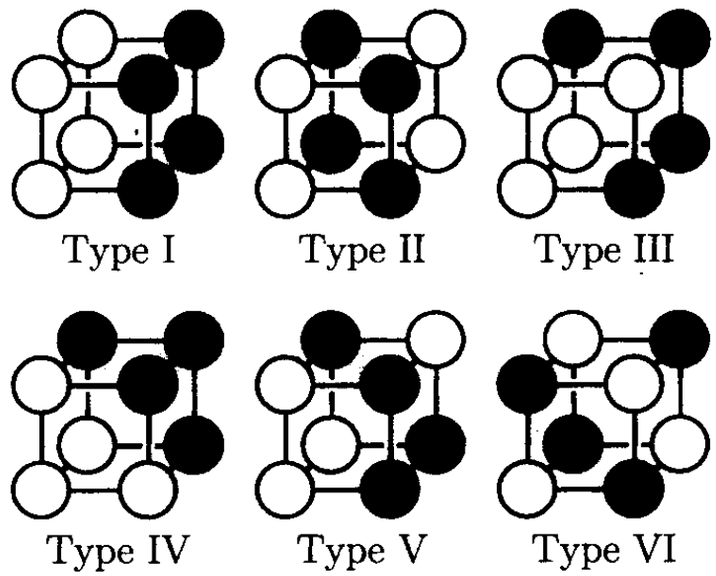
Solutions found for SHJ problems

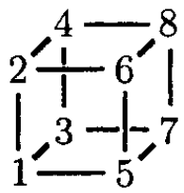
Type I

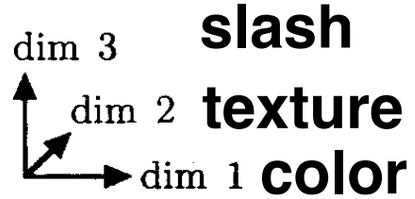
Type II

w_{Ax}







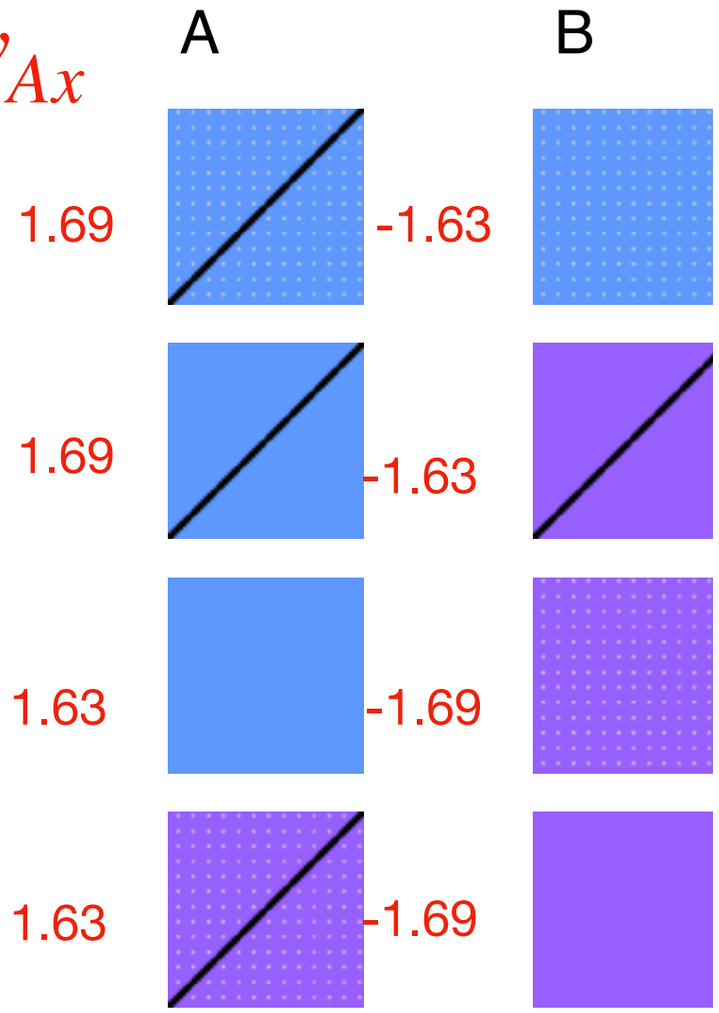


α_{color}	$\alpha_{texture}$	α_{slash}	α_{color}	$\alpha_{texture}$	α_{slash}
0.544	0.0	0.0	0.61	0.0	0.61

Solutions found for SHJ problems

Type III

w_{Ax}

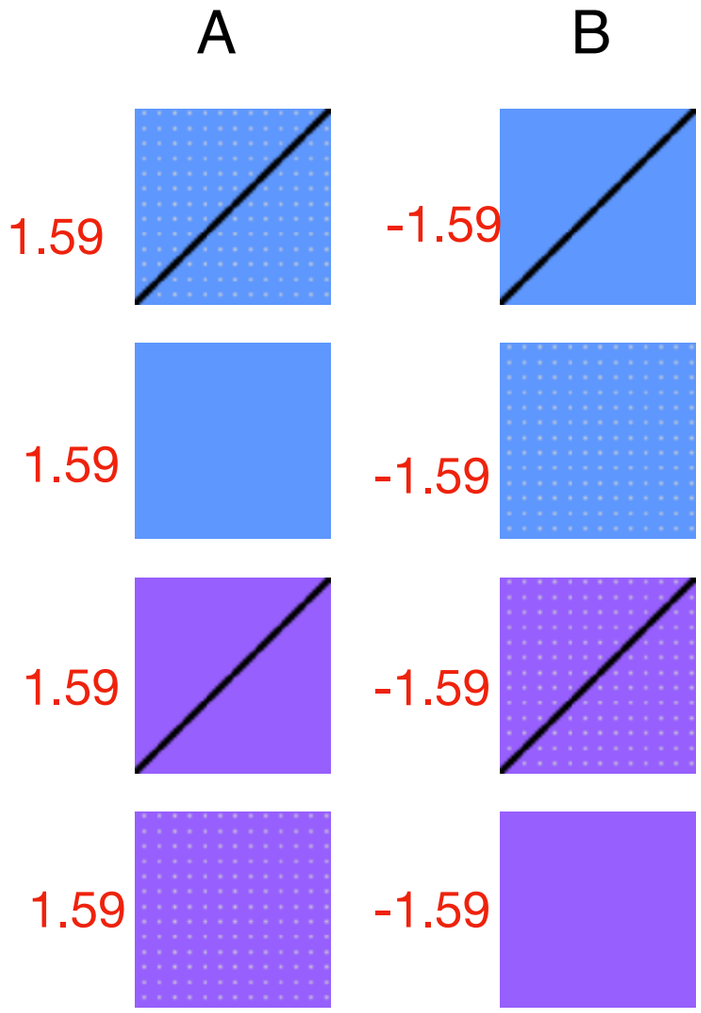


α_{color} 0.399

$\alpha_{texture}$ 0.399

α_{slash} 0.385

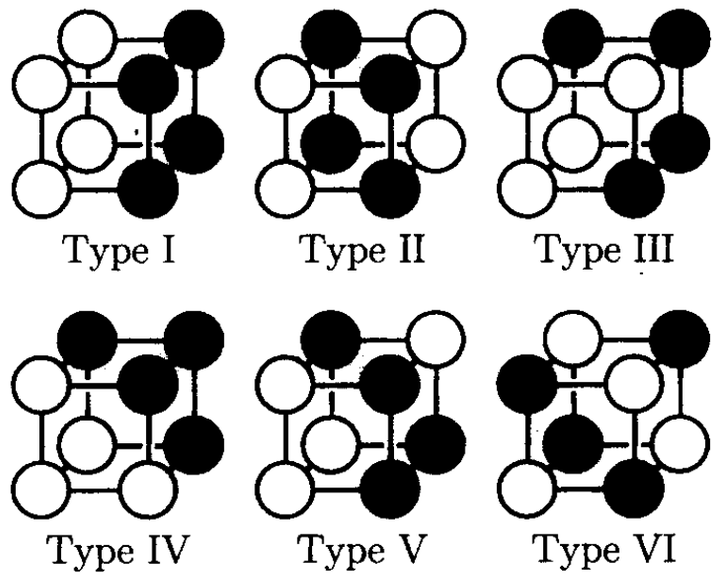
Type VI

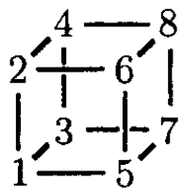


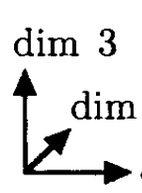
α_{color} 0.62

$\alpha_{texture}$ 0.62

α_{slash} 0.62





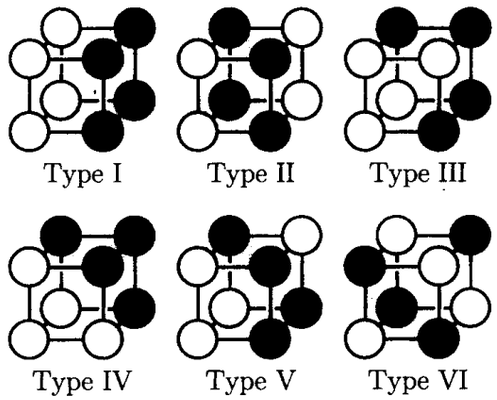


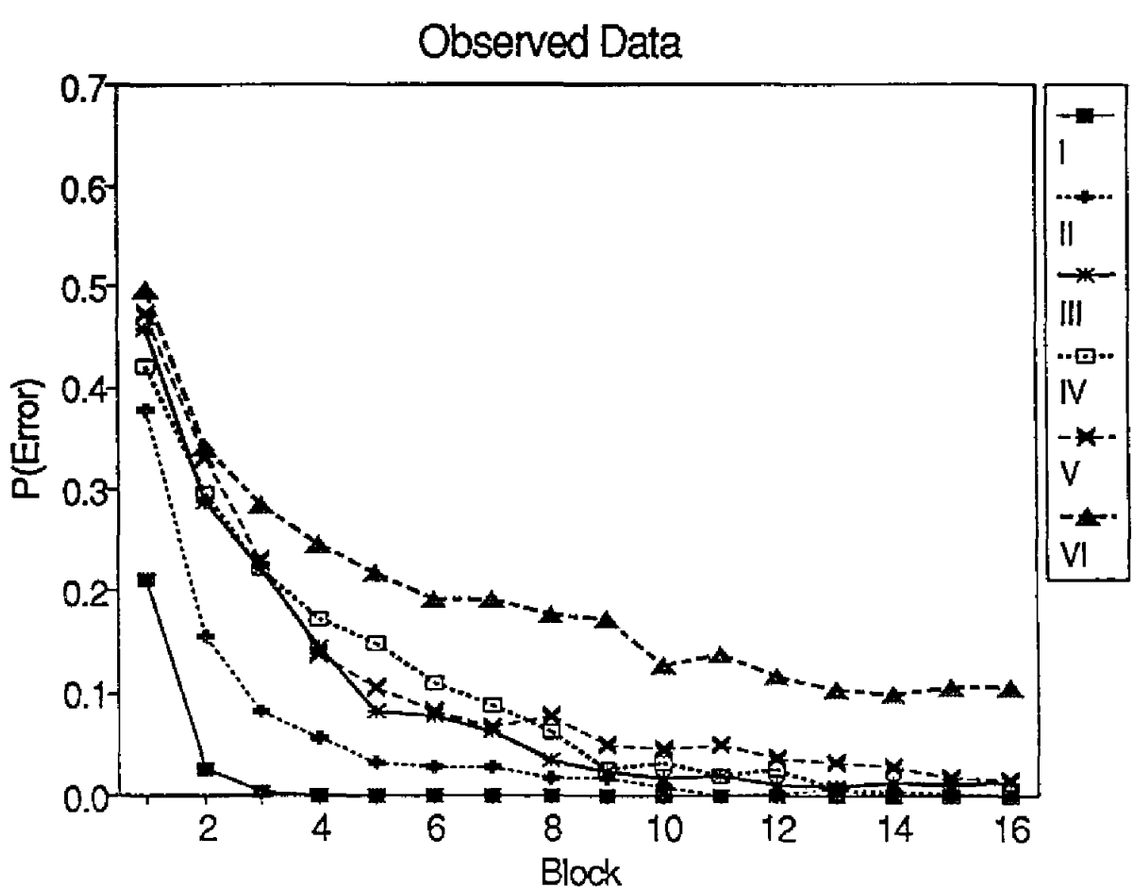
slash

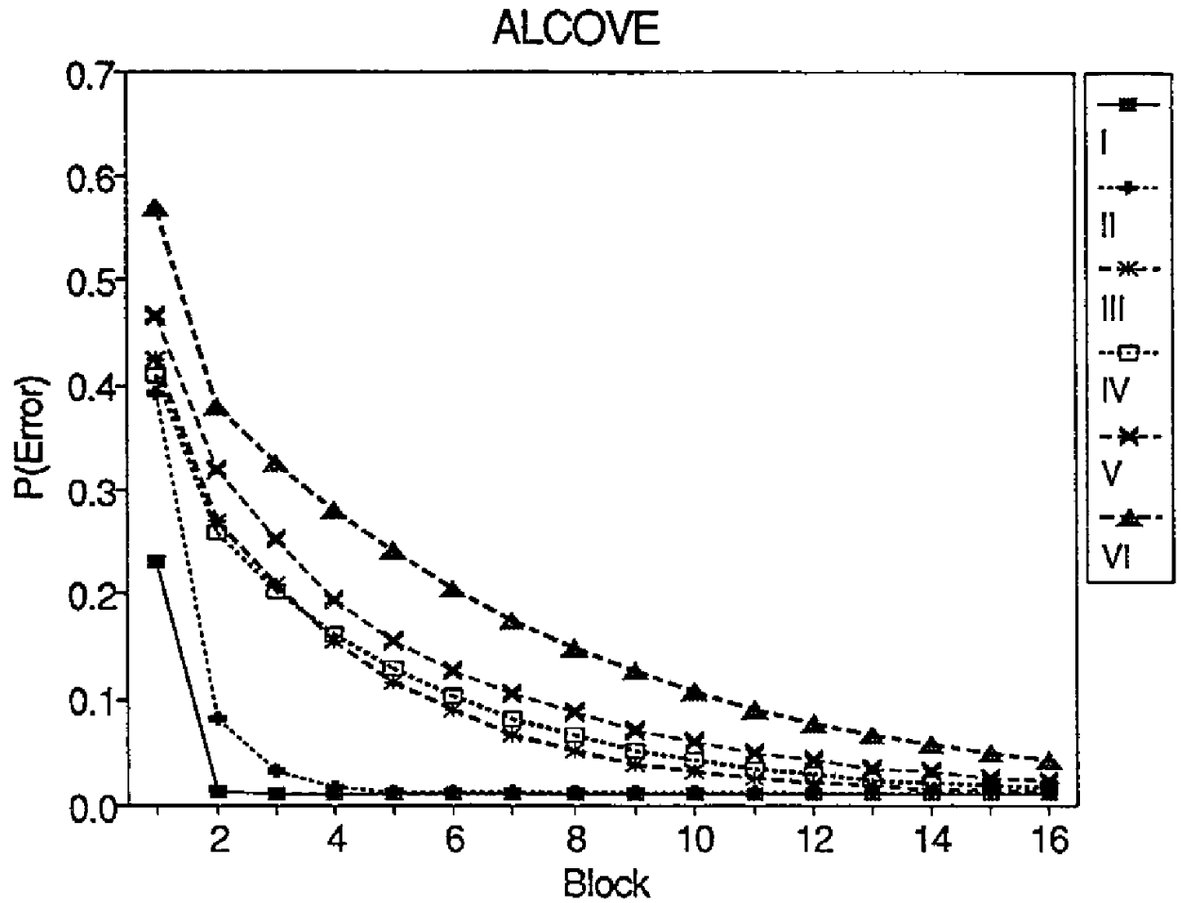
texture

color

Comparing learning curves for people and ALCOVE



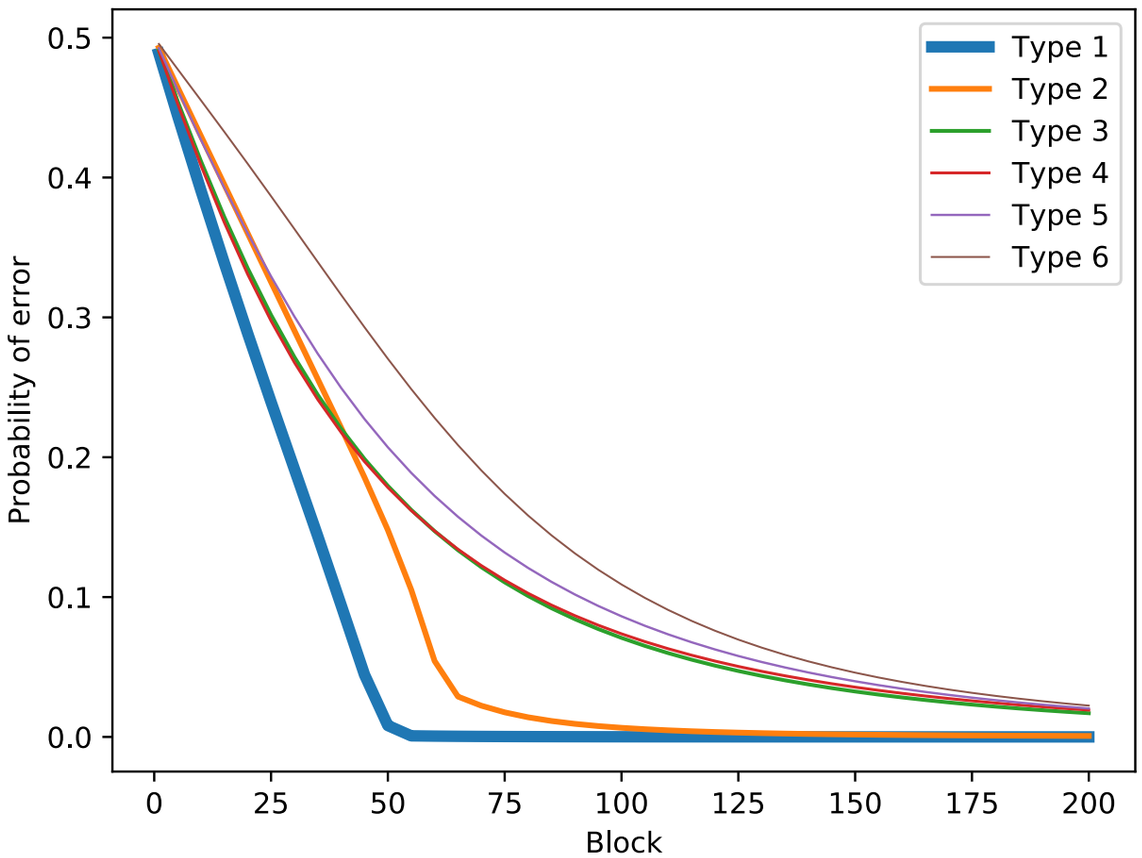




(from Nofosky et al. (1994) SHJ replication in *Memory & Cognition*)

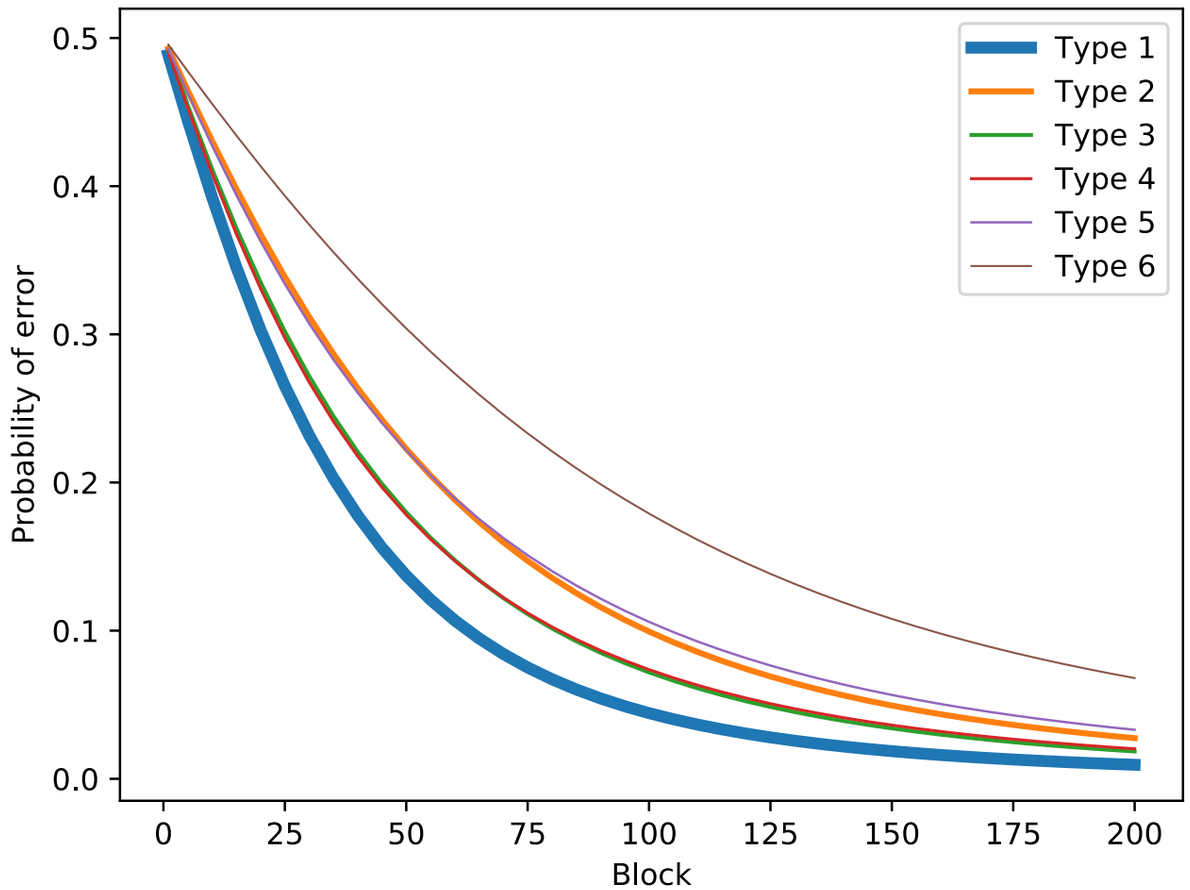
Attentional learning is critical

ALCOVE with attention weights



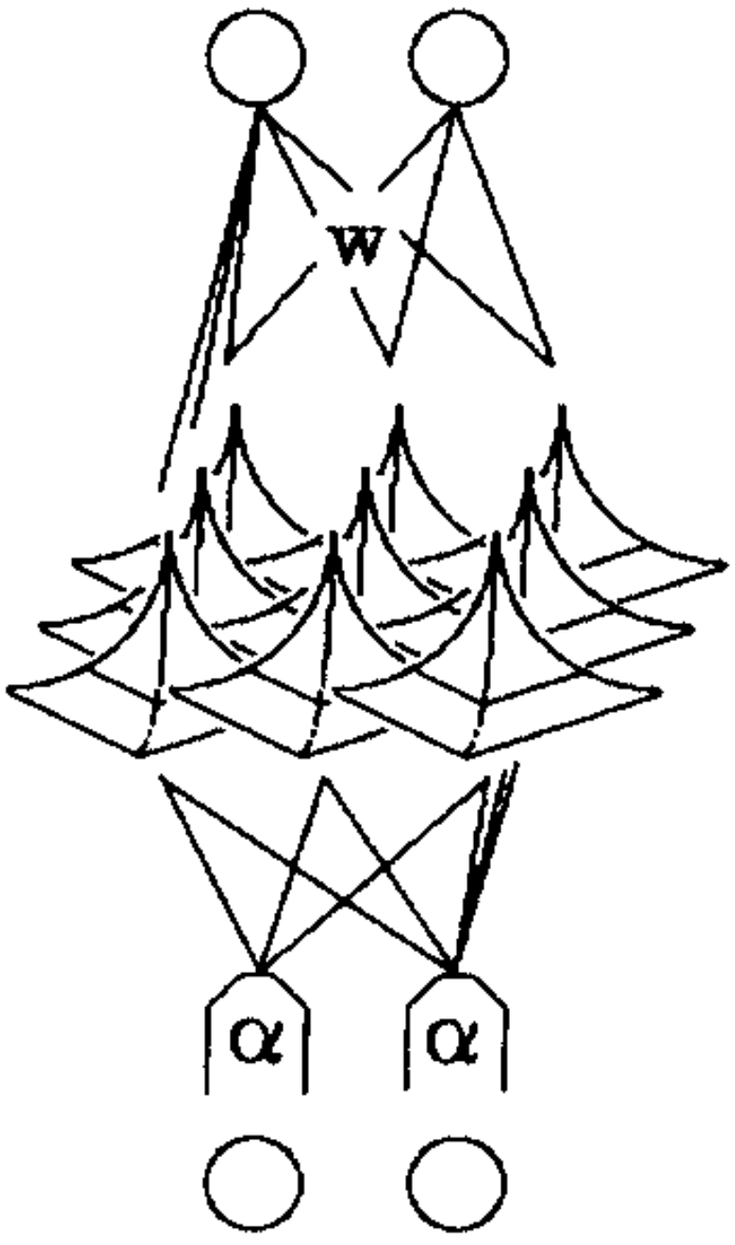
ALCOVE without attention weights

(gets Type II order wrong)

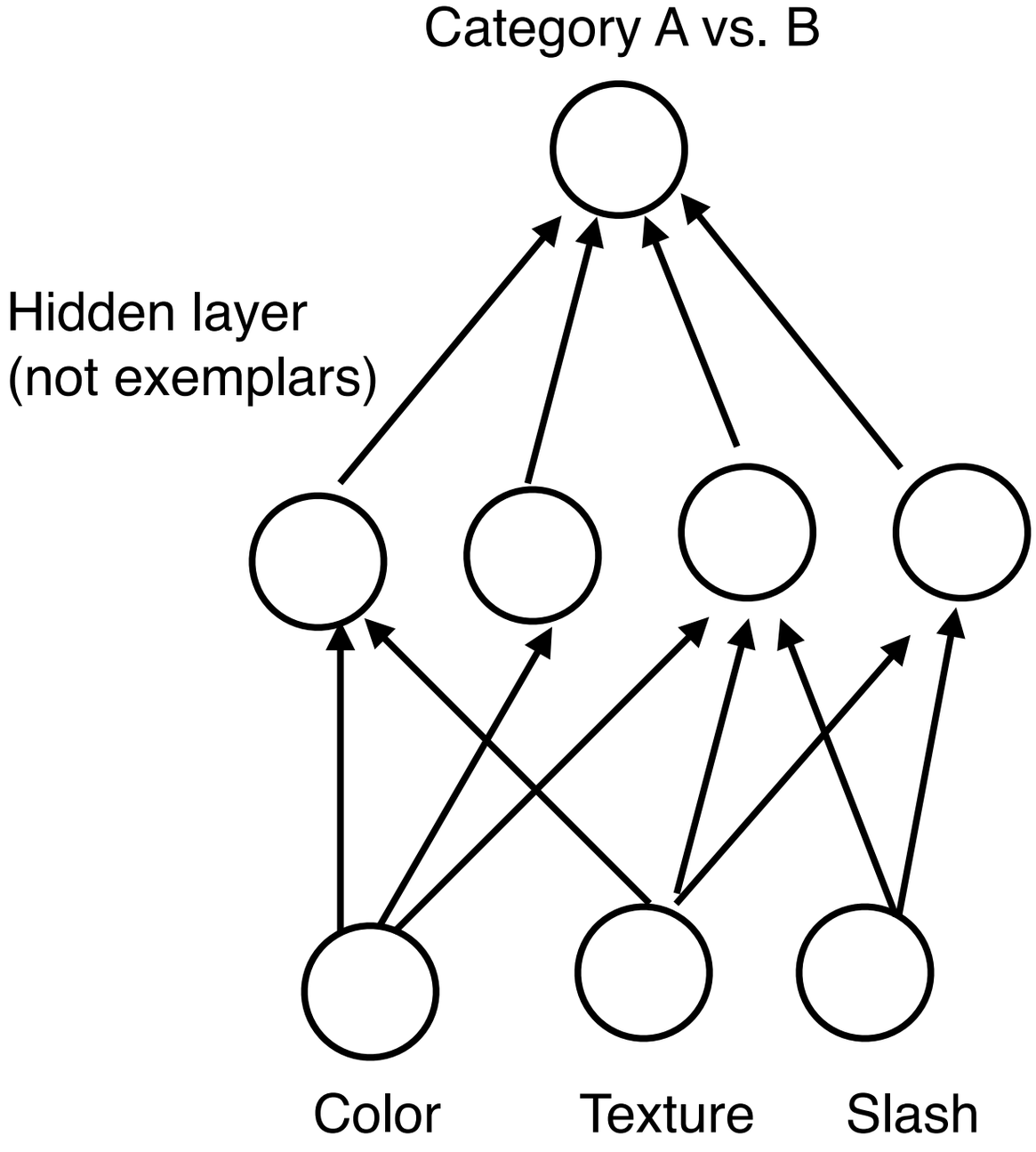


What about a standard neural network?

ALCOVE

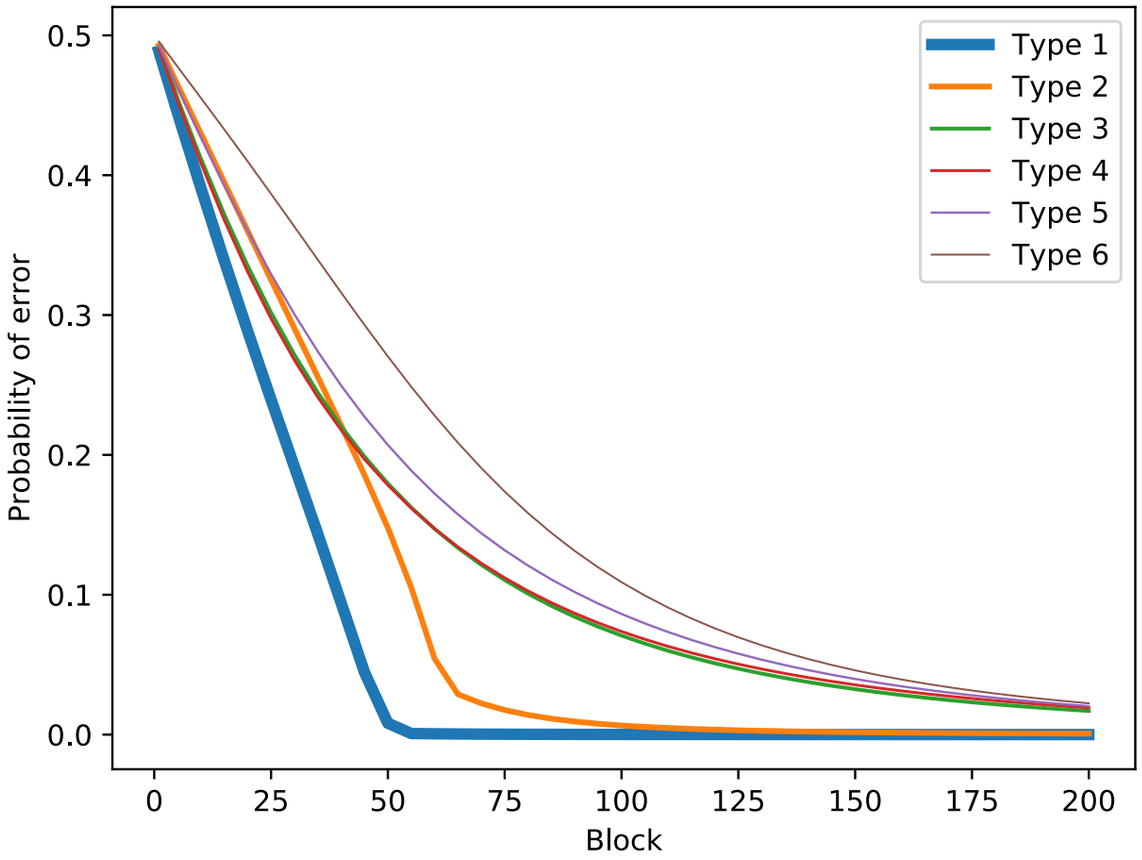


standard multi-layer network



What about a standard neural network?

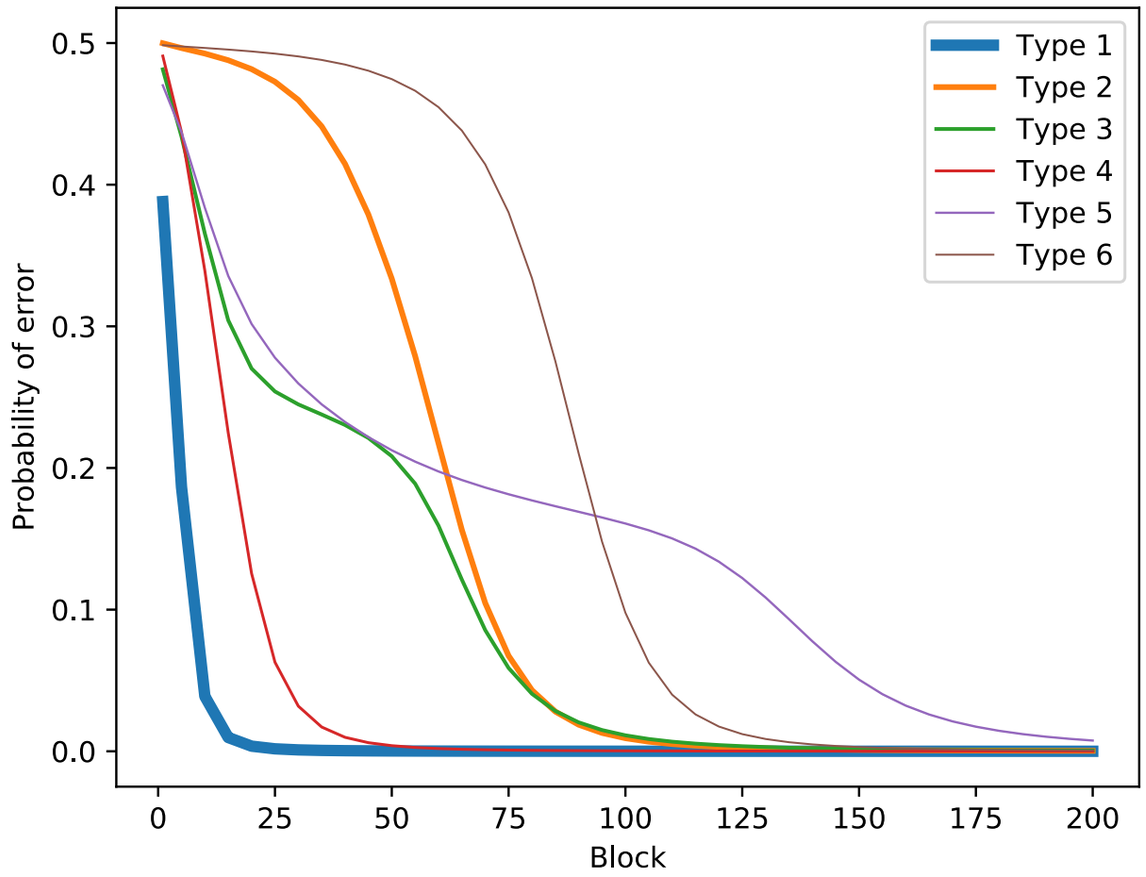
ALCOVE with attention weights



Standard multi-layer neural network

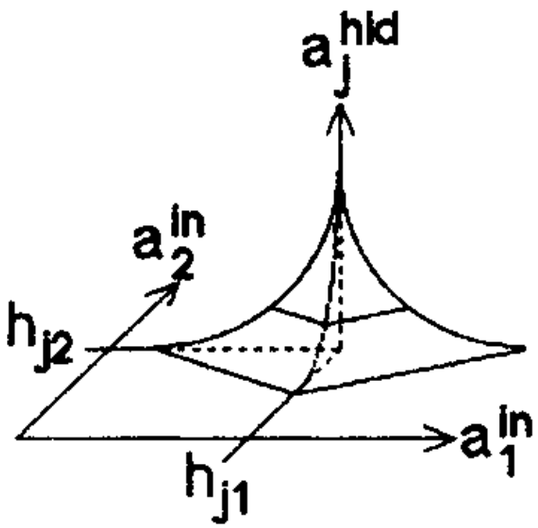
(gets Type II order wrong)

(gets Type IV order very wrong)

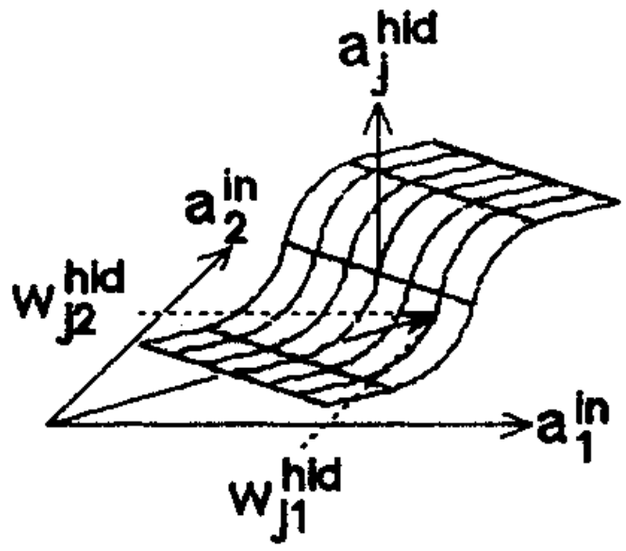


Standard nets prefer linearly separable categories, while ALCOVE prefers strong dimensional alignment

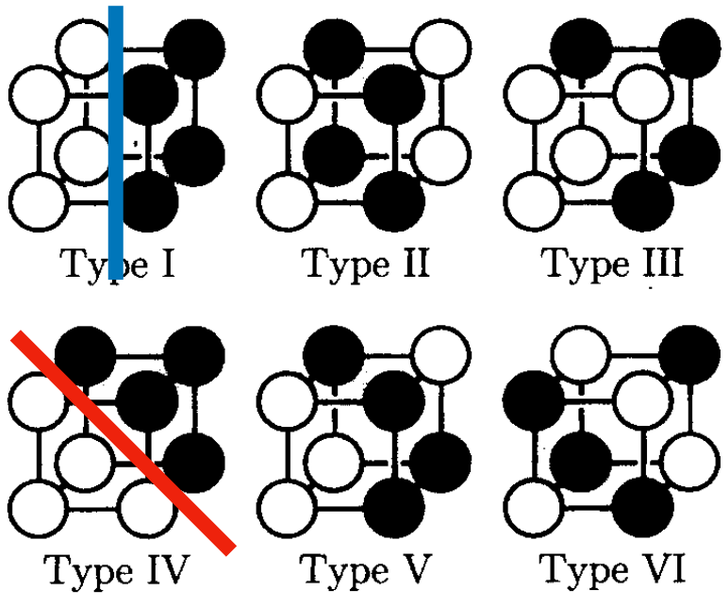
activation profile for ALCOVE hidden unit

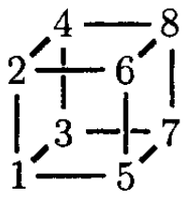


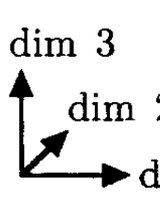
activation profile for standard neural net hidden unit



Standard multi-layer neural network





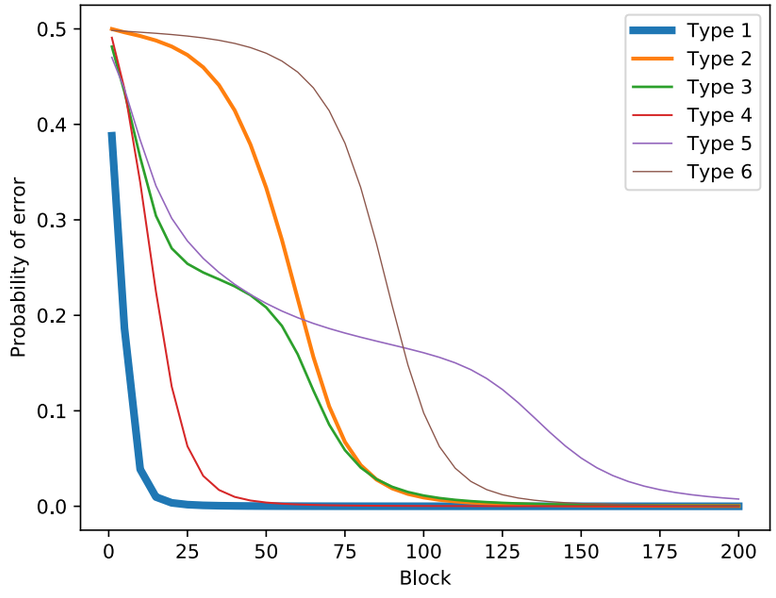


slash

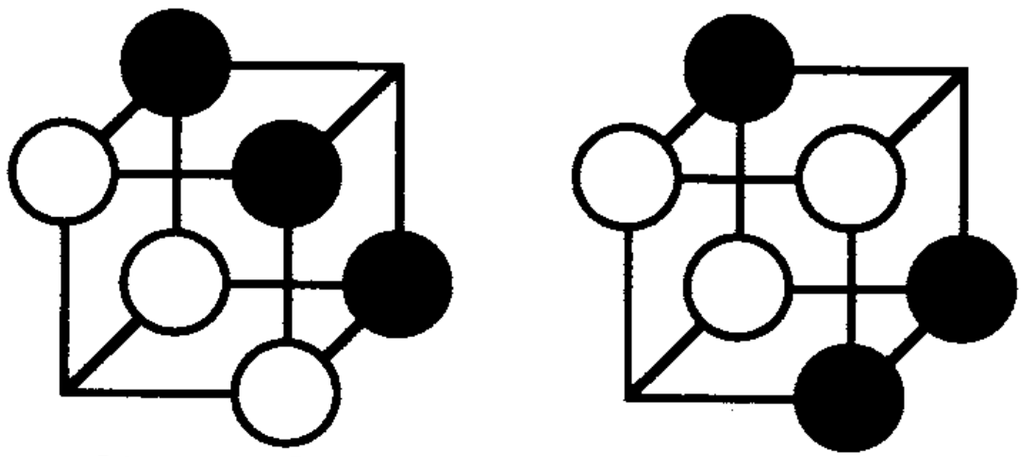
texture

color

Only Type 1 and Type IV are linearly separable, and thus they are learned first with standard net



As predicted by ALCOVE, people don't necessarily favor linearly separable categories



linearly separable

non-linearly separable

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.)

linearly separable condition: people had 39.5 errors on average

non-linearly separable condition: people had 38.0 errors on average

(criterion was two error-free passes through stimuli)

Interlude: Bayes' rule for updating beliefs in light of data

Data (D): John is coughing

Hypotheses:

$h_1 = \text{John has a cold}$

$h_2 = \text{John has emphysema}$

$h_3 = \text{John has a stomach flu}$

"Bayes' rule"

posterior — likelihood — prior

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Which hypotheses should we believe, and with what certainty?

We want to calculate the posterior probabilities: $P(h_1|D)$, $P(h_2|D)$, and $P(h_3|D)$

This example is from Josh Tenenbaum.

Bayesian inference

posterior ← $P(h_i|D)$ = likelihood ← $P(D|h_i)$ prior ← $P(h_i)$ / $\sum_j P(D|h_j)P(h_j)$

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Data (D): John is coughing

Hypotheses:

h_1 = John has a cold

h_2 = John has emphysema

h_3 = John has a stomach flu

$P(h_1) = .75 \quad P(D|h_1) = 1$

$P(h_2) = .05 \quad P(D|h_2) = 1$

$P(h_3) = .2 \quad P(D|h_3) = .2$

Prior favors h_1 and h_3, over h_2

Likelihood favors h_1 and h_2, over h_3

Posterior favors h_1, over h_2 and h_3

$$P(h_1|D) = .89 = \frac{.75(1)}{.75(1) + .05(1) + .2(.2)}$$

$P(h_2|D) = .06$

$P(h_3|D) = .05$

Base rate neglect and ALCOVE

(Experiment from Gluck and Bower 1988)

There is a common disease and rare disease with certain base rates:

$$P(C) = 0.75 \qquad P(R) = 0.25$$

The diseases lead to symptoms with a certain pattern, with s1 arising often in cases of the rare disease:

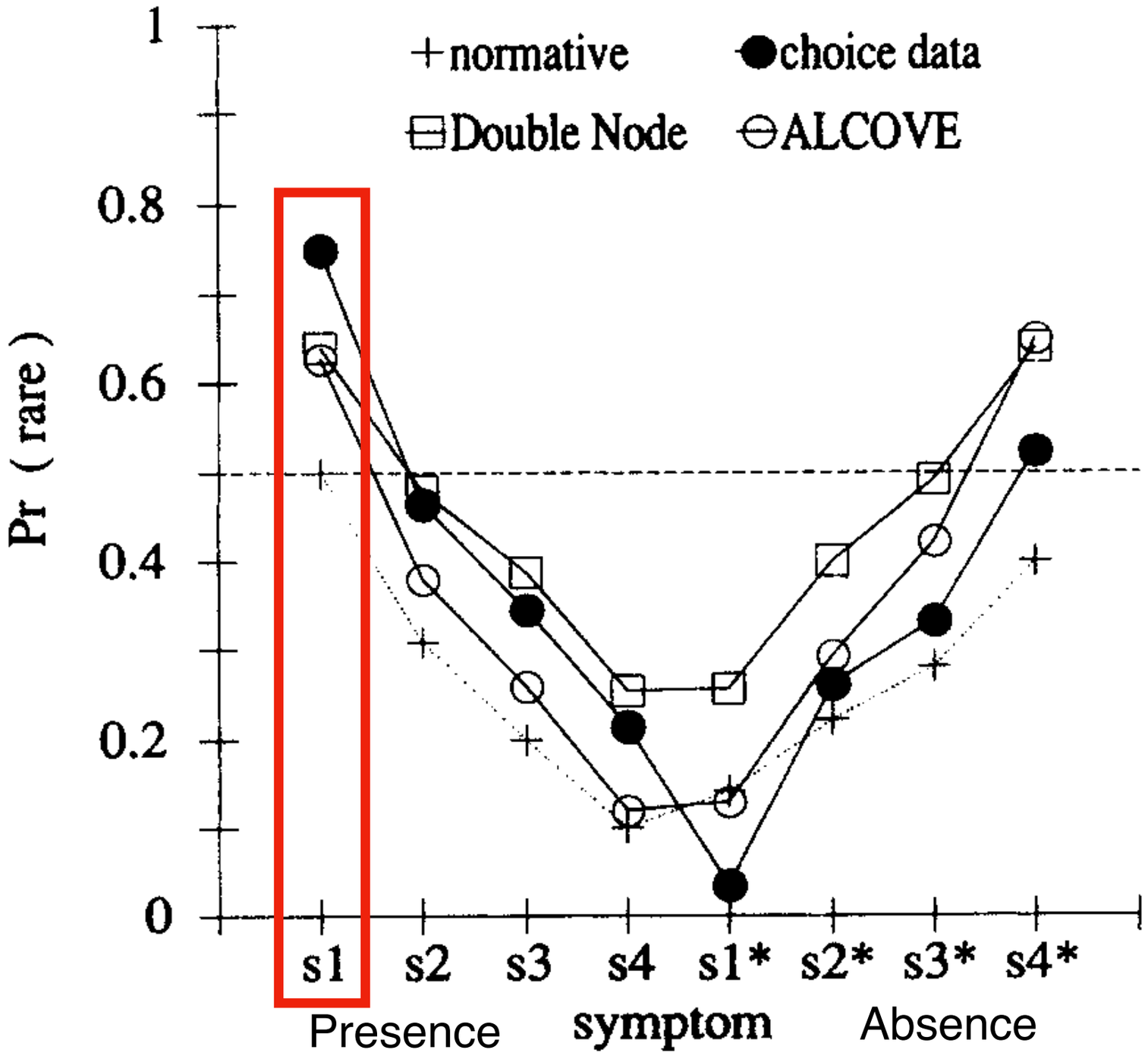
Symptom	Rare disease (R)	Common disease (C)
s1	0.6	0.2
s2	0.4	0.3
s3	0.3	0.4
s4	0.2	0.6

$$P(s1 \mid C) = 0.2$$
$$P(s1 \mid R) = 0.6$$

Both due to mismatch in base rates, the presence of s1 could mean either disease:

$$P(R \mid s1) = \frac{P(s1 \mid R)P(R)}{P(s1 \mid R)P(R) + P(s1 \mid C)P(C)} = 0.5 \qquad P(C \mid s1) = 0.5$$

Base rate neglect



Normative:

$$P(R \mid s1) = \frac{P(s1 \mid R)P(R)}{P(s1)}$$

The context model behaves like the normative model, since it stores all examples with their correct labels

Error driven learning forces exemplars to compete for the right to activate output nodes, resulting in base-rate neglect when s1 is presented in isolation