Psychology 452

Week 5: Perceptrons And Animal Learning

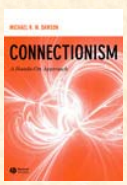
- Mathematical Models
- Mathematical Models Of Conditioning
  - Rescorla-Wagner
  - Logical Neuron
- Equivalence Between Models
  - Empirical
  - Formal
- The Perceptron Paradox

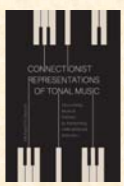
Course Trajectory

When	What
~~Weeks 1-3~~	~~Basics of three architectures (DAM, perceptron, MLP)~~
Weeks 4-6	**Cognitive science of DAMs and perceptrons**
Week 7	Connectionism and Cognitive Psychology
Weeks 8-10	Interpreting MLPs
Weeks 11-13	Case studies (interpretations, applications, architectures)

Discussion?

- Questions, comments or issues?





Models And Psychology

- Simon & Newell (1958) predicted that "within ten years most theories in psychology will take the form of computer programs."
- Models have become important in psychology
  - Models of data
  - Mathematical models
  - Computer simulations
- Let's consider some issues that emerge in mathematical modeling

Herbert Simon (1916-2001)

Allen Newell (1927-1992)

Mathematical Models

"From the first efforts toward psychological measurement, investigators have had in mind the goal of making progress toward generality in psychological theory by developing quantities analogous to mass, charge, and the like in physics and showing that laws and principals formulated in terms of these derive quantities would have greater generality than those formulated in terms of observables" (Estes, 1975)

William K. Estes, Professor Emeritus, Psychology, Indiana University

Properties

Property	Mathematical Models
Analyses of existing data	Yes
Linear transformation	Usually not
Goodness of fit	Yes
Yields surprises	Maybe
Behaves	No

Hull's Law For Growth Of S-R Habits

- "The essential nature of the learning process may, however, be stated quite simply … the process of learning consists in the strengthening of certain of these connections as contrasted with others, or in the setting up of quite new connections" (Hull, 1943).
  - The physiological limit or maximum (M),
  - The ordinal number (N) of the reinforcement producing a given increment to the habit strength ($_SH_R$),
  - The constant factor (f) according to which a portion ($\Delta_S H_R$) of the unrealized potentiality is transferred to the actual habit strength ata given reinforcement.
- Over trials, the law becomes:

$$\Delta_S H_R = f(M - {}_S H_R)$$

Clark Hull

$${}_S^N H_R = M - Me^{-N\log(\frac{1}{1-f})}$$

Pavlovian Conditioning

- Pavlovian conditioning is the process of repeatedly pairing a CS with an US so that eventually the CS will produce this response
- We can gain tremendous insights into this type of learning by considering mathematical equations that attempt to account for it

Ivan Petrovich Pavlov

Before conditioning….

Food (US) ---> Salivation (UCR)

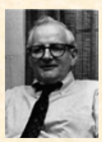
Bell (CS) -----> No Salivation

After conditioning….

Bell (CS) -----> Salivation (CR)

Conditioning And Context

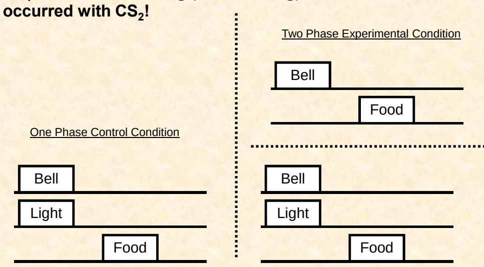
- CS does not occur in a vacuum
- CS appears in the context of numerous other stimuli
- Context, and previous training, can interfere with the desired learning of the CS
- This is nicely demonstrated by Kamin's blocking phenomenon



Leon J. Kamin

Blocking Phenomenon

- If CS_1 (the bell) has already been conditioned to elicit the response, then when it is paired with CS_2 (the light), further learning does not occur
- CS_1 blocks the learning (conditioning) that could have occurred with CS_2!



Explaining Blocking

"Organisms only learn when events violate their expectations. Certain expectations are built up about events following a stimulus context; expectations initiated by the complex and its component stimuli are then only modified when consequent events disagree with the composite expectation" (Rescorla & Wagner, 1972)

Robert Rescorla

Allan Wagner

Rescorla-Wagner Rule

- Rescorla and Wagner used a mathematical model to make their "cognitive" account more rigorous
  - $\Delta V_{(t)}$ - Change in associative strength at time t
  - $V_{(t)}$ - Current associative strength of CS
  - α - Salience of CS
  - λ - Maximum associative strength possible

$$\Delta V_{(t)} = \alpha\left(\lambda - V_{(t)}\right)$$

Multiple CSs

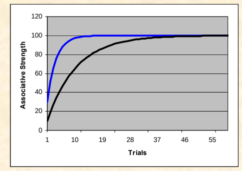
- The Rescorla-Wagner model can easily be generalized to handle situations in which more than one CS can be presented
- This is done by assuming that there is a total associative strength that is the sum of the components, and that the Rescorla-Wagner equation can be selectively applied to each CS

$$\Sigma V = V_A + V_B + V_C$$

$$\Delta V_{A(t)} = \alpha_A \left(\lambda - \Sigma V_{(t)} \right)$$

Formalizing Learning

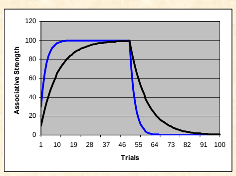
- The Rescorla-Wagner model works by choosing values for the constants, and updating associative strength if the CS is present
- For both runs on the right, $\lambda = 100$
- For the blue line, $\alpha = 0.3$, and for the black line $\alpha = 0.1$



$$\Delta V_{A(t)} = \alpha_A \left(\lambda - \Sigma V_{(t)} \right)$$

Formalizing Extinction

- Extinction is modeled by changing the value of λ to 0 in the weight change equation
- The examples on the right continue the previous example, extinguishing the conditioning from Phase 1



$$\Delta V_{A(t)} = \alpha_A \left(0 - \Sigma V_{(t)} \right)$$

Formalizing CS Absence

- The associative strength is not always modified
- If the CS is not presented, then its associative strength will not change
- The equation below bluntly defines this situation

$$\Delta V_{A(t)} = 0$$

A Unified Rule

- A unified formulation of the Rescorla-Wagner rule accounts for all of its special-case instances that have been shown in previous slides

$$\Delta V_{A(t)} = \alpha_A \left(\lambda \cdot US - \Sigma V_{(t)} \right) \cdot CS_A$$

Learning

$$\begin{aligned} \Delta V_{A(t)} &= \alpha_A \left(\lambda \cdot US - \Sigma V_{(t)} \right) \cdot CS_A \\ &= \alpha_A \left(\lambda \cdot 1 - \Sigma V_{(t)} \right) \cdot 1 \\ &= \alpha_A \left(\lambda - \Sigma V_{(t)} \right) \end{aligned}$$

Extinction

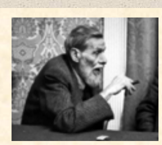
$$\begin{aligned} \Delta V_{A(t)} &= \alpha_A \left(\lambda \cdot US - \Sigma V_{(t)} \right) \cdot CS_A \\ &= \alpha_A \left(\lambda \cdot 0 - \Sigma V_{(t)} \right) \cdot 1 \\ &= \alpha_A \left(0 - \Sigma V_{(t)} \right) \end{aligned}$$

Absent CS

$$\begin{aligned} \Delta V_{A(t)} &= \alpha_A \left(\lambda \cdot US - \Sigma V_{(t)} \right) \cdot CS_A \\ &= \alpha_A \left(\lambda \cdot 1 - \Sigma V_{(t)} \right) \cdot 0 \\ &= 0 \end{aligned}$$

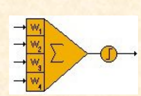
Logical Neurons

- Animal learners have not been the sole developers of formal learning models
- Since the 1940s researchers have formalized neurons, attempting to describe the brain *logically*



Warren McCulloch

Walter Pitts



Formal Neurons Learn

- Some logical neurons, like the ADALINE model developed by Widrow and Hoff in 1959, learn
- Feedback about the correctness of their responses is used to modify connection weights

Bernard Widrow

Marcian Hoff

ADALINE Learning

- ADALINE stood for "adaptive linear element".
- Connections are modified by computing error, which is difference between desired and observed response to a stimulus
- Widrow and Hoff realized ADALINE as a physical system, which is pictured below (the "KNOBBY ADALINE")

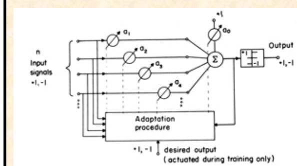


Figure 1. An Automatically-Adapted Threshold Element.

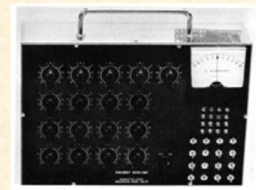
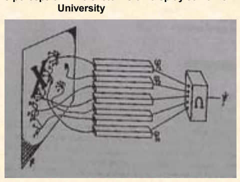


Figure 2. An Elementary Learning Machine.

The Perceptron

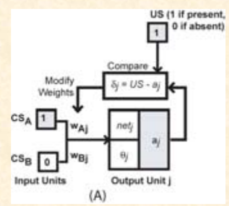
A sketch of Frank Rosenblatt by Peter Kahn, with Minsky and Papert's illustration of his perceptron. The sketch is on display at Cornell University



- The perceptron is an architecture very similar to ADALINE, developed by Frank Rosenblatt in the late 1950s
- It is trained by the delta rule, which is formally equivalent to Widrow-Hoff learning

Perceptron Conditioning With The Delta Rule

- Delta rule learning adjusts weights according to response error, where the response is binary
- If input units are used to represent the presence or absence of CSs, and if the response is viewed as a prediction about the presence of the CS, then you can use a perceptron to model classical conditioning

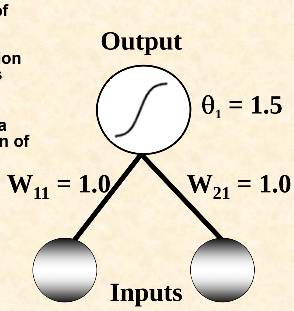


$$\delta_j = t_j - a_j$$

$$\Delta w_{ij} = \eta \cdot a_i \cdot \delta_j$$
$$= \eta \cdot a_i \cdot (t_j - a_j)$$

Integration Device

- Using the terminology of Dana Ballard, call a perceptron that uses a continuous approximation of a step function as its activation function an integration device
- Train this network with a gradient descent version of Rosenblatt's delta rule

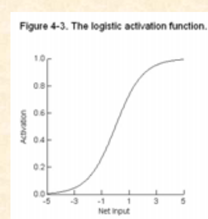


Dana Ballard

The Gradient Descent Rule

- If the derivative of the logistic function is included in the error definition of the delta rule, an integration device can be trained with a gradient descent rule
- The derivative is provided below:

$$f'(net_j) = a_j \cdot (1 - a_j)$$



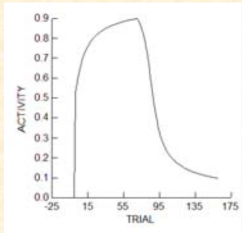
$$\Delta w_{ij} = \eta \cdot a_i \cdot f'(net_j) \cdot \delta_j$$
$$= \eta \cdot a_i \cdot a_j \cdot (1 - a_j)(t_j - a_j)$$

Integration Device Results: Acquisition Curves

- The integration device can be used to simulate many standard paradigms in the animal conditioning literature
- Here we see the learning curve and extinction curve for a simple network trained on a single CS

Phase	Trial Type	Input	t_j
1	A+	1	1
2	A-	1	0

Table 4-2



Integration Device Results: Blocking

- Blocking was historically important in the development of the Rescorla-Wagner rule
- It can be demonstrated in an integration device
- On average, the control networks generate a response of 0.55 to CS_B. In contrast, experimental networks generate an average response of 0.10 to CS_B – functionally equivalent to "off". This difference is statistically significant (*$t = 160.974$, $df = 48$, $p < 0.0001$*).

Blocking Paradigm

Phase	Type	CS_A	CS_B	t_j
1	-	0	0	**0**
	A+	1	0	**1**
2	-	0	0	**0**
	AB+	1	1	**1**

Table 6-5

Weight Changes In A Network

Phase	Unit	Weight	θ
Phase 1	CS_A	4.63	-
	CS_B	0.09	-
	Output	-	-2.20
Phase 2	CS_A	4.64	-
	CS_B	0.09	-
	Output	-	-2.20

Table 6-6

Empirical Equivalences

- Dawson (2008) reports a number of experiments that have shown that integration devices can model many standard conditioning phenomena
  - Classical conditioning of individual stimuli
  - Behaviorally plausible acquisition and extinction curves
  - The effect of CS intensity on the rate of conditioning
  - The effect of US intensity on the rate of conditioning
  - Associations to compound stimuli
  - The discrimination of compound stimuli from their components
  - Overshadowing
  - Blocking
  - Conditioned inhibition
  - Renewal, or context-dependent extinction
  - Superconditioning

Formal Equivalence?

- Sutton and Barto proved the equivalence between a connectionist architecture and a psychological learning rule by translating the Rescorla-Wagner rule into the Widrow-Hoff rule
- A similar proof has been developed by Gluck
- *These proofs assume that the activation function of the output unit is linear!*

Richard Sutton

Andrew Barto

Linear Proof

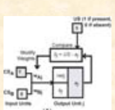
- If activity is identical to net input, and net input is total associative strength, then the error term in the delta rule is identical to the "distance from maximum association term" in the Rescorla-Wagner model
- That is, the Rescorla-Wagner model can be translated into the delta rule under this assumption

$$\Sigma V = V_A + V_B + V_C$$

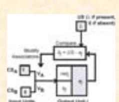
$$net_j = \sum_i a_i w_{ij}$$

$$\Sigma V = \sum_i a_i V_i$$

$$\Sigma V = \sum_i a_i V_i = \sum_i a_i w_{ij} = net_j$$



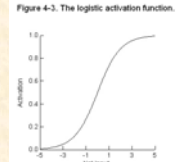
$$a_j = \sum_i a_i w_{ij} = net_j$$



Nonlinear Activity and Association

- Perceptrons and related devices depend on nonlinear activation functions, so the typically cited proof of the relation between animal and machine learning is not really all that applicable
- We need to define how nonlinear output is related to internal associative strength
- One proposal is the equation below:

Figure 4-3. The logistic activation function.



$$\Sigma V = \lambda \cdot f(net_j)$$

Nonlinear Proof

$$\Delta V_{i(t)} = \alpha_i \left(\lambda \cdot US - \Sigma V_{(t)}\right) \cdot CS_i$$

Rescorla-Wagner Rule

$$\Delta V_{i(t)} = \alpha_i \left(\lambda \cdot US - \lambda \cdot f(net_j)_{(t)}\right) \cdot a_i$$

Substitute for total association

$$\Delta V_{i(t)} = \alpha_i \cdot \lambda \cdot \left(US - f(net_j)_{(t)}\right) \cdot a_i$$
$$= \alpha_i \cdot \lambda \cdot \left(US - a_{j(t)}\right) \cdot a_i$$

Move shared terms outside parentheses

$$\Delta V_{i(t)} = \eta \cdot \left(US - a_{j(t)}\right) \cdot a_i$$

Simplify constants

$$\Delta w_{ij} = \eta \cdot \delta_j \cdot a_i$$
$$= \eta \cdot (t_j - a_j) \cdot a_i$$

If US indicates presence, then the equation above is identical to the delta rule, but now for a nonlinear system!

The Perceptron Paradox

- Miller, Barnet and Grahame (1995) have documented many successes and failures of the Rescorla-Wagner model
- Given the formal equivalence that we have established, perceptrons and integration devices must have the same successes and failures
- *The perceptron paradox is that this is not true!*
- We can easily demonstrate learning results in an integration device that diverge from the Rescorla-Wagner model, and in fact improve upon it

Ralph Miller

Robert Barnet

Paradoxical Result: Facilitated Reacquisition

- The Rescorla-Wagner model does not predict facilitated reacquisition after extinction
- This is one of the model's failures
- This failure is not evident in an integration device!
- On average, networks learned the associations to the CSs after 695.9 sweeps during Phase 1. After extinction, networks reacquired these associations after only 62.6 sweeps. Not surprisingly, this difference was statistically significant ($t = 3005.415$, $df = 48$, $p < 0.0001$).

Phase	Trial Type	CS_A	CS_B	t_j
1	-	0	0	**0**
	A+	1	0	**1**
	B+	0	1	**1**
2	-	0	0	**0**
	A-	1	0	**0**
	B-	0	1	**0**
3	-	0	0	**0**
	A+	1	0	**1**
	B+	0	1	**1**

Table 7-1

Paradoxical Result: Extinction of a Conditioned Inhibitor

- The Rescorla-Wagner model predicts that when a conditioned inhibitor is extinguished, its associative strength should become more positive
- This is one of the model's failures, as this prediction is not supported by animal data
- This failure is not evident in an integration device!
- Note the changes in the conditioned inhibitor's weight after phase 2 training

Phase	Trial Type	CS_A	CS_B	t_j
1	-	0	0	0
	A+	1	0	1
	AB-	1	1	0
2	-	0	0	0
	A-	1	0	0
	B-	0	1	0

Table 7-5

Phase	Unit	Weight	θ
1	CS_A	4.47	-
	CS_B	-4.80	-
	Output	-	-2.45
2	CS_A	2.37	-
	CS_B	-4.81	-
	Output	-	-4.58

Table 7-6

Paradoxical Result: Overexpectation

- Let CS_A and CS_B be independently paired with a US. Then, the two CSs are presented as a compound and are paired with the US.
- Overexpectation is defined as occurring when there is reduced responding (relative to a control) to CS_A and CS_B as individual stimuli following the training on the compound stimulus.
- Prediction of this effect was a triumph of the Rescorla-Wagner model
- Dawson and Spetch (2005) argued that the overexpectation effect will not be produced in an integration device, and supported this argument with simulation results

Why Does The Paradox Occur?

- The perceptron paradox arises because integration devices were not just mathematical models of changes in associative strength, but were simulations that had to behave
- Therefore, associative strength must be converted into a response
- The Rescorla-Wagner model is mute with respect to how associative strength becomes behavior

Property	Mathematical Models	Computer Simulations
Analyses of existing data	Yes	Possibly
Linear transformation	Usually not	Usually not
Goodness of fit	Yes	Yes, but nonstandard
Yields surprises	Maybe	Hopefully
Behaves	No	Yes

Implications Of The Paradox

- The Rescorla-Wagner model is evaluated by comparing it to behaving animals
- Such comparisons must involve a tacit theory of how associative strength becomes behavior!
- The nature of this theory has a severe impact on the responses of the model – for instance, you can change the behavioral predictions of the model by changing your theory of behavior, but leaving Rescorla-Wagner untouched!
- For example, Dawson and Spetch (2005) have shown that changing the activation function has implications for whether overexpectation is observed. This change does not affect a "Rescorla-Wagner" account of machine learning!!

Figure 4-4. The Gaussian activation function.