An Introduction to Learning

Lecture 7/13

Todd M. Gureckis, Lila Davachi
Department of Psychology
New York University
{todd.gureckis, lila.davachi}@nyu.edu
Agenda for Today

- Generalization and the Representation of Similarity
  - Basic concepts in generalization and discrimination
  - Shepard’s Universal Law of Generalization
  - Possible ways of representing similarities in the brain
  - Negative patterning and configural unit models
  - Neural substrates of learning and generalization
1 Basic concepts in generalization and discrimination
Generalization and Discrimination

- Effective behavior in the world requires organisms to generalize from past experience in appropriate ways.
  - Advantages include savings in amount of direct experience, ability to avoid stimuli \textit{without} trying them out/sampling, handling novel situations and making predictions about what to expect, guiding efficient search of the environment.

- In addition, animals must learn to discriminate between objects or situations that (although appear most similar) lead to different outcomes.

- An inherent trade-off, generalization often implies a lack of discrimination and vice versa.

- Aka “stimulus control of operant behavior?!” (remember, Skinner box has that light!)
Simple Forms of Generalization and Discrimination

Animals will generalize their experience to new stimuli (e.g., from a 800 Mhz tone to a 850 Mhz tone)

The shape of this drop-off in responding is known at the generalization gradient

Roughly exponentially shaped with a peak in the center (similar for all types of stimuli: colors, tones, shapes, textures, angles)

**Figure 6-4.** Pavlovian Generalization Gradients of Excitation (A) with a Tone and Inhibition (B) with a Light as CS+. In Part A, maximum responding occurs to the CS+ and drops off systematically as the tones presented are more different from the CS+. In Part B, minimum responding occurs to the CS− responding increases as lights are more and more different from the CS−.
Question: is it just a failure to be able to tell apart particular colors/tones?
Answer: No. Can train pretty fine discrimination between things.

Reshaping of generalization gradient with training.
Learning and Generalization in Networks

(a) Train "yellow"

Output nodes

Modifiable weights

Input nodes

Response

(b) Test "yellow-orange"

No response

Green  Yellow-green  Yellow  Yellow-orange  Orange

Green  Yellow-green  Yellow  Yellow-orange  Orange
Learning and Generalization in Networks
Learning and Generalization in Networks

(a) Distributed representation network

(b) Train "yellow"

(c) Test "yellow-orange": some decline in responding

(d) Test "orange": more decline in responding

- Output node
- Modifiable weights
- Internal representation nodes

- Input nodes
- Modifiable weights
- Internal representation nodes
Learning and Generalization in Networks

(a) Tone
Response
Output nodes
1
Modifiable weights
Tone
Input nodes
1

(b) Light
Response
Output nodes
1
Modifiable weights
Light
Input nodes
1

(c) Tone + light
Response
Output nodes
2
Modifiable weights
Tone Light
Input nodes
1 1
Learning and Generalization in Networks

(a) Tone alone

Response

Output nodes
Modifiable weights

Tone only Tone + light Light only

1 0 0

1 0 0

1 0 0

Input nodes
Fixed weights

Tone Light

1 0

1 0

1 0

(b) Light alone

Response

No response

Output nodes
Modifiable weights

Tone only Tone + light Light only

1 1 1

1 1 1

1 1 1

Input nodes
Fixed weights

Tone Light

1 0

1 0

1 0

(c) Tone + light

Response

Output nodes
Modifiable weights

Tone only Tone + light Light only

1 1 1

1 1 1

1 1 1

Input nodes
Fixed weights

Tone Light

1 0

1 0

1 0
What is the basis for generalization?

It seems that key to explaining how organisms generalize from one situation to the next is some notion of *similarity*.

However, similarity in physical structure of the world doesn’t seem to be enough, particularly in light of the fact that generalization behavior for some types of physical stimuli look quite different:
What is the basis for generalization?

“At midcentury, influential behavioral scientists [including Lashley, Bush, & Mosteller] were reaching the discouraging conclusion that there could be no invariant law of generalization. If we took physical difference as the independent variable, gradients of generalization, reflecting properties of the particular animal as much as the physically measured differences between the stimuli, could not be expected to be uniform or even monotonic. If, instead, we sough a psychological measure of difference as the independent variable, the most basic data would be the generalization data themselves--apparently rendering the attempt to determine a functional law entirely circular.” (pg. 1318)
What is the basis for generalization?

“Instead of starting with a physical parameter space, I proposed to start with the generalization data and to ask: Is there an invariant monotonic function whose inverse will uniquely transform those data into numbers interpretable as distances in some appropriate metric space?” (pg. 1318)

Distance in a psychological space

What is the basis for generalization?

What is the space?

MDS (Multi-dimensional Scaling)

Given a set of rated similarities between stimuli (e.g., item-item confusions), figure out a N-dimensions space to place the items in which the euclidean distances in this space can be related to the similarities through a non-parametric monotonic relationship (aka non-metric MDS, Shepard & Kruskal, 1964)
How? Conceptually simple, computationally challenging

1. Choose number of dimensions N

2. Place each point in random location in space

3. Compute distance between pairs of points in the space (euclidean most often)

3. Use various types of regression (including non-parametric) to see if distances are rank ordered correctly. In perfect case, rank order of distance in space would match the rank order of similarities in original matrix (a score about goodness of fit is provided by a “stress” measure)

4. Continue to improve configuration using gradient descent or some other search process until find a good fit.
\[ p_{ij} = \text{probability that a response learned to stimulus } i \text{ is made to stimulus } j \]

\[ g_{ij} = \text{generalization from stimulus } i \text{ to } j \]

(normalized)
\[ g_{ij} = \sqrt{p_{ij} \times p_{ji} / p_{ii} \times p_{jj}} \]


Plot distance in MDS space \( d_{ij} \) compared to empirical values \( g_{ij} \) for a variety of stimuli across 12 experiments.

$p_{ij}$ = probability that a response learned to stimulus $i$ is made to stimulus $j$

$g_{ij}$ = generalization from stimulus $i$ to $j$ (normalized)

$g_{ij} = \sqrt{p_{ij} \times p_{ji} / p_{ii} \times p_{jj}}$

Plot distance in MDS space ($d_{ij}$) compared to empirical values ($g_{ij}$) for a variety of stimuli across 12 experiments.

Wow.
$p_{ij} =$ probability that a response learned to stimulus $i$ is made to stimulus $j$

$g_{ij} =$ generalization from stimulus $i$ to $j$ (normalized)

$$g_{ij} = \sqrt{p_{ij} \times p_{ji} / p_{ii} \times p_{jj}}$$

Plot distance in MDS space ($d_{ij}$) compared to empirical values ($g_{ij}$) for a variety of stimuli across 12 experiments


How? What is the space?

stimulus $i$ to $j$

(normalized)

$$g_{ij} = \sqrt{p_{ij} \times p_{ji} / p_{ii} \times p_{jj}}$$

Plot distance in MDS space ($d_{ij}$) compared to empirical values ($g_{ij}$) for a variety of stimuli across 12 experiments
Colors

Tones separated by octaves

“Only in relation to such abstract spatial representations can we achieve an invariant law” (pg. 1319)
What about the distance metric?

**Minkowski norm**

\[ d_{ij} = \left( \sum_{k=1}^{K} |x_{ik} - x_{jk}|^r \right)^{1/r} \]

**Euclidean metric**

\[ r = 2 \]

**City block metric**

\[ r = 1 \]

**Shapes varying in size and orientation**

**Colors**
Why exponential?

**Fig. 2.** (A) A centrally symmetric convex region shown as centered on 0, as centered on \( x \), and as having a center, \( c \), falling within the intersection of the regions centered on 0 and on \( x \). (B) For an illustrative nonconvex region centered on 0, the locus of centers, \( c \), of similarly shaped regions having a constant (approximately 20\%) overlap with the region centered on 0 (dotted curve); and an ellipse corresponding to the Euclidean metric (smooth curve).

Consequential Region
Why exponential?

\[ p(y \in C | x) = \sum_{h : y \in h} p(h | x) \]

\[ p(h | x) \]

\[ h \in H \]
Why exponential?

The 1-D Case

\[ g(x) = \int_0^\infty p(s) \frac{m(s,x)}{m(s)} ds \]  \hspace{1cm} (2)

Because the size of the consequential region cannot be negative and is assumed to have finite expectation \( \mu \), \( p(s) \) is zero for all \( s < 0 \), and (in addition to being nonnegative itself) satisfies the two conditions

\[ \int_0^\infty p(s) ds = 1 \]  \hspace{1cm} (3)

\[ \int_0^\infty s \cdot p(s) ds = \mu < \infty \]  \hspace{1cm} (4)
Why exponential?

The 2-D Case

For stimuli, like colors, that differ along dimensions that do not correspond to uniquely defined independent variables in the world, moreover, psychological space should have no preferred axes. The consequential region is then most reasonably assumed to be circular or, whatever other shapes may be assumed, to have all possible orientations in the space with equal probability. Symmetry then entails strictly circular contours of equal generalization (Fig. 4B) and, hence, the Euclidean metric (or $L_2$-norm).

For stimuli that differ along dimensions, such as size and orientation, that correspond to uniquely defined independent variables in the world, however, psychological space should possess, corresponding preferred axes. Whatever type of shape is then assumed for the consequential region, the degree to which that region is extended along one preferred axis should not be correlated with the degree to which it is extended along another such axis. Instead of assuming that the region is a square or circle, in the two-dimensional case, the individual might assume that it is a rectangle or an ellipse aligned with the preferred axes of the space. Integration must then be carried out over the two independently variable size dimensions of the consequential region, say $s$ and $t$ (as indicated on the right in Fig. 4), with corresponding probability density functions, $p(s)$ and $p(t)$. 
Why exponential?

The 2-D Case

For stimuli, like colors, that differ along dimensions that do not correspond to uniquely defined independent variables in the world, moreover, psychological space should have no preferred axes. The consequential region is then most reasonably assumed to be circular or, whatever other shapes may be assumed, to have all possible orientations in the space with equal probability. Symmetry then entails strictly circular contours of equal generalization (Fig. 4B) and, hence, the Euclidean metric (or $L_2$-norm).

For stimuli that differ along dimensions, such as size and orientation, that correspond to uniquely defined independent variables in the world, however, psychological space should possess, corresponding preferred axes. Whatever type of shape is then assumed for the consequential region, the degree to which that region is extended along one preferred axis should not be correlated with the degree to which it is extended along another such axis. Instead of assuming that the region is a square or circle, in the two-dimensional case, the individual might assume that it is a rectangle or an ellipse aligned with the preferred axes of the space. Integration must then be carried out over the two independently variable size dimensions of the consequential region, say $s$ and $t$ (as indicated on the right in Fig. 4), with corresponding probability density functions, $p(s)$ and $p(t)$.

\[ d_{ij} = \left( \sum_{k=1}^{K} |x_{ik} - x_{jk}|^r \right)^{1/r} \]

Minkowski norm!
Neural reinstatement

credit: Alexa Tompary
Neural reinstatement

credit: Alexa Tompary

O’Reilly & McClelland 1994
Cortical reactivation

Hippocampus?

credit: Alexa Tompary

Wheeler et al 2000

Evidence for neural reinstatement as a function of memory

Predictions

Hippocampus
Perirhinal cortex (PrC)

credit: Alexa Tompary

Specificity of reinstatement

Does neural reinstatement reflect general or specific episodes?

Encoding

Retrieval

Same cue
Same associate

Different cue
Same associate

Pearson correlation

credit: Alexa Tompary
Representation similarity analysis

- Trial 1: Same context
- Trial 2: Same context
- Project-specific methods
- Details
- Other stuff

credit: Alexa Tompary
## Representation similarity analysis

**Trial 1**

<table>
<thead>
<tr>
<th>1.6</th>
<th>0.9</th>
<th>1.5</th>
<th>2.1</th>
<th>2.8</th>
<th>1.1</th>
<th>0.3</th>
<th>1.3</th>
<th>0.1</th>
</tr>
</thead>
</table>

Euclidean distance from trial 1: \(d = 0.33\)

**Trial 2**

<table>
<thead>
<tr>
<th>1.6</th>
<th>0.9</th>
<th>1.5</th>
<th>2.0</th>
<th>2.7</th>
<th>1.2</th>
<th>0.1</th>
<th>1.3</th>
<th>0.3</th>
</tr>
</thead>
</table>

\(d = 0.33\)

**Trial 3**

<table>
<thead>
<tr>
<th>2.5</th>
<th>1.9</th>
<th>2.1</th>
<th>0.1</th>
<th>2.7</th>
<th>0.5</th>
<th>1.1</th>
<th>1.3</th>
<th>0.3</th>
</tr>
</thead>
</table>

\(d = 2.68\)

Credit: Alexa Tompary
Correlations with trial 1

- Trial 1: $r=0.89$
- Trial 2: $r=0.45$

Similarity matrix

credit: Alexa Tompary
Decoding

<table>
<thead>
<tr>
<th>Image</th>
<th>fMRI scan</th>
<th>Voxel pattern</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="shoe.png" alt="Shoe" /></td>
<td><img src="brain.png" alt="Brain" /></td>
<td><img src="voxel.png" alt="Voxel" /></td>
<td>=SHOE</td>
</tr>
<tr>
<td><img src="cat.png" alt="Cat" /></td>
<td><img src="brain.png" alt="Brain" /></td>
<td><img src="voxel.png" alt="Voxel" /></td>
<td>=CAT</td>
</tr>
<tr>
<td><img src="shoe.png" alt="Shoe" /></td>
<td><img src="brain.png" alt="Brain" /></td>
<td><img src="voxel.png" alt="Voxel" /></td>
<td>=SHOE?</td>
</tr>
</tbody>
</table>

During testing, the program must guess the object viewed on the basis of what it has learned about similar patterns of activity.

credit: Nature
Generalization Data
(similarity rating stimulus confusions)

Psychological Space
(Multidimensional space representing arrangements of stimuli)

Neural Space
Pattern similarity in terms of voxels
The need for biases

- 1 million marbles in the urn
- You can draw 100 marbles out
- Is it possible to determine which numbers occur at least once among the millions in the urn?
The need for biases

- 1 million marbles in the urn
- You can draw 100 marbles out
- Is it possible to determine which numbers occur at least once among the millions in the urn?

Answer: Of course, no. Every marble could be different
The need for biases

- 1 million marbles in the urn
- You can draw 100 marbles out
- Is it possible to determine which numbers occur at least once among the millions in the urn?

**Answer:** Yes, if all the same number!
The need for biases

Answer: Of course, no. Every marble could be different

Intermediate cases?
What if we assume only five types of marbles in urn?

Answer: Yes, if all the same number!
The need for biases

If totally unbiased generalization systems are incapable of making the inductive leap to characterize the new instances, then the power of a generalization system follows directly from its biases – from decisions based on criteria other than consistency with the training instances. Therefore, progress toward understanding learning mechanisms depends upon understanding the sources of, and justification for, various biases.

Mitchell (1980)

Possible biases

- Domain knowledge
- Intended use/goal of generalization (e.g., cost of being incorrect... i.e., risk sensitive)
- Knowledge about the source of training data
- Biases towards simplicity/generality
- Analogy with previous generalizations
Generalization, similarity, and Bayesian inference

\[ p(h \mid x) \]

\[ h \in H \]

\[ p(y \in C \mid x) = \sum_{h : y \in h} p(h \mid x) \]

\[ p(h \mid x) = \frac{p(x \mid h)p(h)}{p(x)} \]

\[ = \frac{p(x \mid h)p(h)}{\sum_{h' \in H} p(x \mid h')p(h')} \]
Strong vs. Weak Sampling

- What is learned will depend on the learners *assumptions* about the situation.

- Where the data generated from the true concept or were they generated at random (independent of the concept?)

- This map onto the notion of STRONG versus WEAK sampling

\[
p(x|h) = \begin{cases} 
1 & \text{if } x \in h \\
0 & \text{otherwise} 
\end{cases} \quad \text{[weak sampling]}
\]

\[
p(x|h) = \begin{cases} 
\frac{1}{|h|} & \text{if } x \in h \\
0 & \text{otherwise} 
\end{cases} \quad \text{[strong sampling]}
\]
What is the consequence?
What is the consequence?

Figure 2: Performance of three concept learning algorithms on the rectangle task.
What is the consequence?
Interim summary

- Generalization behavior depends on **inductive biases** (see Mitchell article)

- This is the only thing that can constraint which of the infinite set of generalizations are possible in any particular case

- A related issue is the sampling assumptions that the learner brings to the task... inferences from strongly sampled data can be strong then inferences from weakly sampled data

- This can greatly speed learning and increase confidence in true hypothesis.

- When is strong or weak sampling justified?
Take-home point

- Generalization is not the inverse of failures of discrimination

- The probability of generalization approximates an exponential decay function of distance in an appropriately defined psychological space (but this is a latent variable that has to be studied in and of itself... a focus on “cognitive”/”mental” events)

- To the degree that the shape of consequential stimuli along dimension of that space are correlated or uncorrelated, psychological distances in that space will approximate either the Euclidean or non-Euclidean metrics

- These two facts appear to be a natural implication of a type of “probabilistic geometry”
... but

- Is feature similarity enough to explain the “psychological space” of similarity?
Next time

Memory!

Pavlov

Ebbinghaus
Readings for next time

Coming soon. Check the website Friday afternoon.

Exam posted by Friday
References for Slides


