Revealing priors on category structures through iterated learning

Tom Griffiths
University of California, Berkeley

Brian Christian
Brown University

Mike Kalish
University of Louisiana, Lafayette

Inductive biases

Many of the questions studied in cognitive science involve *inductive problems*, where people evaluate underdetermined hypotheses using sparse data.

Examples:

Learning languages from utterances

blicket toma
dax wug
blicket wug

S → X Y
X → {blicket,dax}
Y → {toma, wug}

Learning functions from (x,y) pairs

Learning categories from instances of their members

Solving inductive problems requires *inductive biases*: *a priori* preferences that make it possible to choose among hypotheses. These biases limit the hypotheses that people entertain, and determine how much evidence is needed to accept a particular hypothesis.

Examples:

Compositional vs. holistic languages
Linear vs. non-linear functions
Categories defined by one-dimensional vs. multidimensional rules

Understanding how people solve inductive problems requires understanding their inductive biases.

Bayesian inference

A framework for stating rational solutions to inductive problems, in which inductive biases are made explicit.

Posterior probability — Likelihood — Prior probability

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h' \in H} P(d \mid h')P(h')}$$

Sum over space of hypotheses

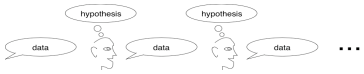
h: hypothesis
d: data

Inductive biases are encoded in the prior distribution.

How can we discover the priors of human learners?

In this work, we develop a novel method for revealing the priors of human learners, and test this method using stimuli for which people's inductive biases are well understood - category structures.

Iterated learning



Much of human knowledge is not learned from the world directly, but from other people (e.g. language).

Kirby (2001) calls this process *iterated learning*, with each learner generating the data from which the next learner forms a hypothesis.

When the learners are Bayesian agents, choosing hypotheses by sampling from their posterior distribution, the probability that a learner chooses a particular hypothesis converges to the prior probability of that hypothesis as iterated learning proceeds.

(Griffiths & Kalish, 2005)

By reproducing iterated learning in the laboratory, can we discover the nature of human inductive biases?

Iterated concept learning

data **hypotheses**

Each learner sees examples from a species of amoebae, and identifies the other members of that species (with a total of four amoebae per species).

Iterated learning is run within-subjects, since the predictions are the same as for between-subjects. The hypothesis chosen on one trial is used to generate the data seen on the next trial, with the new amoebae being selected randomly from the chosen species.

Bayesian model

(Tenenbaum, 1999; Tenenbaum & Griffiths, 2001)

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h' \in H} P(d \mid h')P(h')}$$

d: *m* amoebae
h: |*h*| amoebae

$$P(d \mid h) = \begin{cases} 1/|h|^m & d \in h \\ 0 & \text{otherwise} \end{cases}$$

$$P(h \mid d) = \frac{P(h)}{\sum_{h' | d \in h'} P(h')}$$

Posterior is renormalized prior

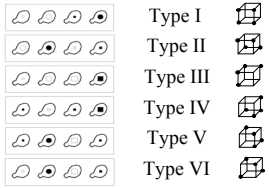
What is the prior?

Category structures

(Shepard, Hovland, & Jenkins, 1961)

Three binary features and four objects per category results in 70 possible category structures.

Collapsing over negations and feature values reduces this to six types of structure.

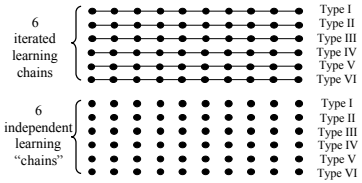


Design and Analysis

Six iterated learning chains were run, each started with a category structure of one of the six types, with subsequent structures being determined by the responses of the participants.

As a control, six "independent" chains were run at the same time, with a structure of the appropriate type being generated randomly at each generation.

With a total of 10 iterations per chain, trials were divided into 10 blocks of 12, with the order of the chains randomized within blocks.

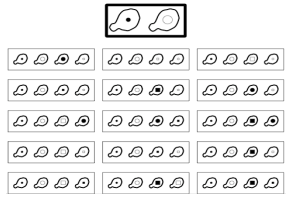


For each experiment, the prior probability assigned to each of the six types of structures was estimated at the same time as classifying participants into two groups: those that responded in a way that was consistent with the prior, and those that selected randomly among the possible structures (consistent with a uniform prior). This was done using the Expectation-Maximization (EM) algorithm. The responses of the participants classified as non-random were then analyzed further.

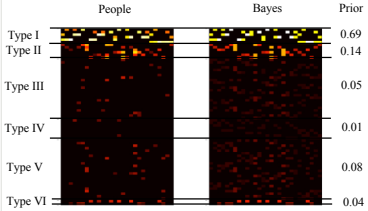
Two experiments examined convergence to the prior and how well the dynamics of iterated learning were predicted by the Bayesian model.

Experiment 1: Two examples

A total of 117 participants performed an iterated concept learning task where they saw two examples from a category, and had to guess the remainder.

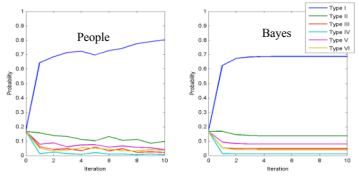


Estimating the prior

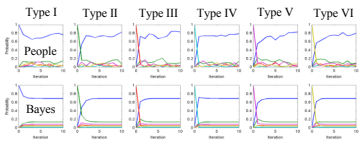


The prior was estimated from the choices of hypotheses in both the iterated learning and independent trials.

Results (n = 69)



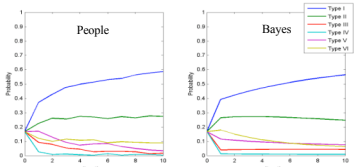
Convergence to the prior occurred rapidly, as emphasized by the results for the iterated learning chains started with different structures.



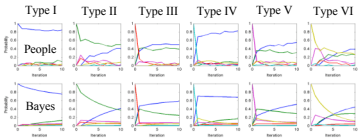
Experiment 2: Three examples

A total of 73 participants performed an iterated concept learning task where they saw three examples from a category, and had to guess the remainder.

Results (n = 64)



Convergence to the prior was slower, as predicted by the Bayesian model. The iterated learning chains started with different structures now exhibited distinctive dynamics mirrored in the human data.



Conclusions

Iterated learning may provide a valuable experimental method for investigating human inductive biases.

With stimuli for which inductive biases are well understood - simple category structures - iterated learning converges to a distribution consistent with those biases.

The dynamics of iterated learning correspond closely with the predictions of a Bayesian model.

Future work will explore what this method can reveal about inductive biases for other kinds of hypotheses, such as languages and functions.