Revealing inductive biases through iterated learning

Tom Griffiths
Department of Psychology
Cognitive Science Program
UC Berkeley

with Mike Kalish, Brian Christian, Simon Kirby, Mike Dowman, Stephan Lewandowsky
Inductive problems

Learning languages from utterances

bicket toma \quad S \rightarrow X \ Y

dax wug \quad X \rightarrow \{bicket,dax\}

bicket wug \quad Y \rightarrow \{toma, \ wug\}

Learning categories from instances of their members

Learning functions from \((x,y)\) pairs
Learning

hypothesis

data
hypothesis

data
Iterated learning
(Kirby, 2001)

What are the consequences of learners learning from other learners?
Outline

Part I: Formal analysis of iterated learning

Part II: Iterated learning in the lab
Outline

Part I: Formal analysis of iterated learning

Part II: Iterated learning in the lab
Objects of iterated learning

- Knowledge communicated across generations through provision of data by learners

- Examples:
  - religious concepts
  - social norms
  - myths and legends
  - causal theories
  - language
Language

• The languages spoken by humans are typically viewed as the result of two factors
  – individual learning
  – innate constraints (biological evolution)

• This limits the possible explanations for different kinds of linguistic phenomena
Linguistic universals

- Human languages possess universal properties
  - e.g. compositionality
    (Comrie, 1981; Greenberg, 1963; Hawkins, 1988)

- Traditional explanation:
  - linguistic universals reflect strong innate constraints specific to a system for acquiring language
    (e.g., Chomsky, 1965)
Cultural evolution

• Languages are also subject to change via cultural evolution (through iterated learning)

• Alternative explanation:
  – linguistic universals emerge as the result of the fact that language is learned anew by each generation (using general-purpose learning mechanisms, expressing weak constraints on languages)
  
  (e.g., Briscoe, 1998; Kirby, 2001)
Analyzing iterated learning

\[ P_L(h|d) \]: probability of inferring hypothesis \( h \) from data \( d \)

\[ P_P(d|h) \]: probability of generating data \( d \) from hypothesis \( h \)
Markov chains

- Variables $x^{(t+1)}$ independent of history given $x^{(t)}$
- Converges to a stationary distribution under easily checked conditions (i.e., if it is ergodic)

Transition matrix
$T = P(x^{(t+1)}|x^{(t)})$
Analyzing iterated learning

\[
d_0 \xrightarrow{P_L(h|d)} h_1 \xrightarrow{P_p(d|h)} d_1 \xrightarrow{P_L(h|d)} h_2 \xrightarrow{P_p(d|h)} d_2 \xrightarrow{P_L(h|d)} h_3 \xrightarrow{} \]

A Markov chain on hypotheses

\[
h_1 \xrightarrow{\Sigma_d P_p(d|h) P_L(h|d)} h_2 \xrightarrow{\Sigma_d P_p(d|h) P_L(h|d)} h_3 \xrightarrow{} \]

A Markov chain on data

\[
d_0 \xrightarrow{\Sigma_h P_L(h|d) P_p(d|h)} d_1 \xrightarrow{\Sigma_h P_L(h|d) P_p(d|h)} d_2 \xrightarrow{\Sigma_h P_L(h|d) P_F} \]
Bayesian inference

Reverend Thomas Bayes
Bayes’ theorem

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

Posterior probability
Likelihood
Prior probability

\( h \): hypothesis
\( d \): data

Sum over space of hypotheses
A note on hypotheses and priors

• No commitment to the nature of hypotheses
  – discrete parameters (Gibson & Wexler, 1994)

• Priors do not necessarily represent innate constraints specific to language acquisition
  – not innate: can reflect independent sources of data
  – not specific: general-purpose learning algorithms also have inductive biases expressible as priors
Iterated Bayesian learning

Assume learners sample from their posterior distribution:

\[
P_L(h \mid d) = \frac{P_P(d \mid h) P(h)}{\sum_{h' \in H} P_P(d \mid h') P(h')}
\]
Stationary distributions

• Markov chain on $h$ converges to the prior, $P(h)$

• Markov chain on $d$ converges to the “prior predictive distribution”

\[ P(d) = \sum_h P(d \mid h)P(h) \]

(Griffiths & Kalish, 2005)
Explaining convergence to the prior

• Intuitively: data acts once, prior many times
• Formally: iterated learning with Bayesian agents is a *Gibbs sampler* on $P(d,h)$

(Griffiths & Kalish, in press)
Gibbs sampling

For variables $\mathbf{x} = x_1, x_2, \ldots, x_n$

Draw $x_i^{(t+1)}$ from $P(x_i|x_{-i})$

$x_{-i} = x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)}$

Converges to $P(x_1, x_2, \ldots, x_n)$

(Geman & Geman, 1984)

(a.k.a. the heat bath algorithm in statistical physics)
Gibbs sampling

(MacKay, 2003)
Explaining convergence to the prior

When target distribution is $P(d,h) = P_P(d|h)P(h)$, conditional distributions are $P_L(h|d)$ and $P_P(d|h)$
Implications for linguistic universals

• When learners sample from $P(h|d)$, the distribution over languages converges to the prior
  – identifies a one-to-one correspondence between inductive biases and linguistic universals
Iterated Bayesian learning

\[ P_L(h \mid d) = \frac{P_P(d \mid h)P(h)}{\sum_{h' \in H} P_P(d \mid h')P(h')} \]

Assume learners sample from their posterior distribution:
From sampling to maximizing

\[ P_L(h \mid d) \propto \left[ \sum_{h' \in H} \frac{P_P(d \mid h)P(h)}{\sum_{h' \in H} P_P(d \mid h')P(h')} \right]^r \]
From sampling to maximizing

• General analytic results are hard to obtain
  – \( r = \infty \) is Monte Carlo EM with a single sample

• For certain classes of languages, it is possible to show that the stationary distribution gives each hypothesis \( h \) probability proportional to \( P(h)^r \)
  – the ordering identified by the prior is preserved, but not the corresponding probabilities

(Kirby, Dowman, & Griffiths, in press)
Implications for linguistic universals

• When learners sample from $P(h|d)$, the distribution over languages converges to the prior
  – identifies a one-to-one correspondence between inductive biases and linguistic universals

• As learners move towards maximizing, the influence of the prior is exaggerated
  – weak biases can produce strong universals
  – cultural evolution is a viable alternative to traditional explanations for linguistic universals
Analyzing iterated learning

• The outcome of iterated learning is strongly affected by the inductive biases of the learners
  – hypotheses with high prior probability ultimately appear with high probability in the population
• Clarifies the connection between constraints on language learning and linguistic universals…
• …and provides formal justification for the idea that culture reflects the structure of the mind
Outline

Part I: Formal analysis of iterated learning

Part II: Iterated learning in the lab
Inductive problems

Learning languages from utterances

blicket toma \[S \rightarrow X \ Y\]
dax wug \[X \rightarrow \{blicket, dax\}\]
blicket wug \[Y \rightarrow \{toma, wug\}\]

Learning categories from instances of their members

Learning functions from \((x, y)\) pairs
Revealing inductive biases

• Many problems in cognitive science can be formulated as problems of induction
  – learning languages, concepts, and causal relations
• Such problems are not solvable without bias
  (e.g., Goodman, 1955; Kearns & Vazirani, 1994; Vapnik, 1995)
• What biases guide human inductive inferences?

If iterated learning converges to the prior, then it may provide a method for investigating biases
Serial reproduction
(Bartlett, 1932)

- Participants see stimuli, then reproduce them from memory
- Reproductions of one participant are stimuli for the next
- Stimuli were interesting, rather than controlled
  – e.g., “War of the Ghosts”
General strategy

• Use well-studied and simple stimuli for which people’s inductive biases are known
  – function learning
  – concept learning

• Examine dynamics of iterated learning
  – convergence to state reflecting biases
  – predictable path to convergence
Iterated function learning

- Each learner sees a set of \((x,y)\) pairs
- Makes predictions of \(y\) for new \(x\) values
- Predictions are data for the next learner

(Kalish, Griffiths, & Lewandowsky, in press)
Function learning experiments

Examine iterated learning with different initial data
<table>
<thead>
<tr>
<th>Initial data</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table represents an initial data set with iterations from 1 to 9.
Iterated concept learning

- Each learner sees examples from a species
- Identifies species of four amoebae
- Iterated learning is run within-subjects

(Griffiths, Christian, & Kalish, 2006)
Two positive examples

data ($d$)
hypotheses ($h$)
Bayesian model

(Tenenbaum, 1999; Tenenbaum & Griffiths, 2001)

\[ P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h' \in H} P(d \mid h')P(h')} \quad d: 2 \text{ amoebae} \]

\[ h: \text{ set of 4 amoebae} \]

\[ P(d \mid h) = \begin{cases} 1/|h|^m & d \in h \\ 0 & \text{otherwise} \end{cases} \quad m: \# \text{ of amoebae in the set } d (= 2) \]

\[ |h|: \# \text{ of amoebae in the set } h (= 4) \]

\[ P(h \mid d) = \frac{P(h)}{\sum_{h' \mid d \in h'} P(h')} \quad \text{Posterior is renormalized prior} \]

What is the prior?
Classes of concepts
(Shepard, Hovland, & Jenkins, 1961)

Class 1

Class 2

Class 3

Class 4

Class 5

Class 6
Experiment design (for each subject)

6 iterated learning chains

6 independent learning “chains”

Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Estimating the prior
Estimating the prior

Human subjects

Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Two positive examples

\((n = 20)\)
Two positive examples
\((n = 20)\)

Human learners

Bayesian model
Three positive examples

data ($d$)

hypotheses ($h$)
Three positive examples

\((n = 20)\)
Three positive examples

\( (n = 20) \)

Human learners

Bayesian model
Identifying inductive biases

• Formal analysis suggests that iterated learning provides a way to determine inductive biases

• Experiments with human learners support this idea
  – when stimuli for which biases are well understood are used, those biases are revealed by iterated learning

• What do inductive biases look like in other cases?
  – continuous categories
  – causal structure
  – word learning
  – language learning
Conclusions

• Iterated learning provides a lens for magnifying the inductive biases of learners
  – small effects for individuals are big effects for groups
• When cognition affects culture, studying groups can give us better insight into individuals
The “information bottleneck”  
(Kirby, 2001)

size indicates compressibility

“survival of the most compressible”
Discovering the biases of models

Generic neural network:
Discovering the biases of models

EXAM (Delosh, Busemeyer, & McDaniel, 1997):
Discovering the biases of models

POLE (Kalish, Lewandowsky, & Kruschke, 2004):
A simple language model

- "agents"
  - 0
  - 1

- "nouns"
  - 0
  - 1

- "verbs"
  - 0
  - 1

- "actions"
  - 0
  - 1

- Events
- Language

- Utterances
A simple language model

- Data: $m$ event-utterance pairs
- Hypotheses: languages, with error $\varepsilon$

\[ p(h) \]
\[ \frac{\alpha}{4} \]
\[ \frac{(1 - \alpha)}{256} \]
Convergence to priors

\( \alpha = 0.50, \varepsilon = 0.05, m = 3 \)

\( \alpha = 0.01, \varepsilon = 0.05, m = 3 \)

Compositionality emerges only when favored by the prior.
The information bottleneck

$\alpha = 0.50, \varepsilon = 0.05, m = 1$

$\alpha = 0.50, \varepsilon = 0.05, m = 3$

$\alpha = 0.50, \varepsilon = 0.05, m = 10$

No effect of amount of data
An example: Gaussians

• If we assume…
  – data, \(d\), is a single real number, \(x\)
  – hypotheses, \(h\), are means of a Gaussian, \(\mu\)
  – prior, \(p(\mu)\), is Gaussian(\(\mu_0, \sigma_0^2\))
• …then \(p(x_{n+1}|x_n)\) is Gaussian(\(\mu_n, \sigma_x^2 + \sigma_n^2\))

\[
\mu_n = \frac{x_n / \sigma_x^2 + \mu_0 / \sigma_0^2}{1 / \sigma_x^2 + 1 / \sigma_0^2} \quad \sigma_n^2 = \frac{1}{1 / \sigma_x^2 + 1 / \sigma_0^2}
\]
An example: Gaussians

- If we assume...
  - data, \( d \), is a single real number, \( x \)
  - hypotheses, \( h \), are means of a Gaussian, \( \mu \)
  - prior, \( p(\mu) \), is Gaussian(\( \mu_0, \sigma_0^2 \))
- \( p(x_n+1|x_n) \) is Gaussian(\( \mu_n, \sigma_x^2 + \sigma_n^2 \))
- \( p(x_n|x_0) \) is Gaussian(\( \mu_0 + c^n x_0, (\sigma_x^2 + \sigma_0^2)(1 - c^{2n}) \))
  i.e. geometric convergence to prior
  \[
  c = \frac{1}{1 + \frac{\sigma_x^2}{\sigma_0^2}}
  \]
\[ \mu_0 = 0, \; \sigma_0^2 = 1, \; x_0 = 20 \]

Iterated learning results in rapid convergence to prior
An example: Linear regression

• Assume
  – data, \( d \), are pairs of real numbers \((x, y)\)
  – hypotheses, \( h \), are functions

• An example: linear regression
  – hypotheses have slope \( \theta \) and pass through origin
  – \( p(\theta) \) is Gaussian\((\theta_0, \sigma_0^2)\)
\[ \theta_0 = 1, \quad \sigma_0^2 = 0.1, \quad y_0 = -1 \]