A probabilistic model of theory formation

Charles Kemp, Joshua B. Tenenbaum, Sourabh Niyogi, Thomas L. Griffiths

Cognition 114 (2010) 165–196

Seminar E, SS 10
Three main questions

- What are theories?
- How do they support inductive inference?
- How are they acquired?
What are theories?

Theories in mathematical logic
Specifying a definite non-empty conceptual class $E$, the elements of which are called elementary statements, a theory $T$ is a conceptual class consisting of certain of these elementary statements that are said to be true.

Deductive theories
The content of $T$ is based on some formal deductive system (formal language and inference rules). Some of its elementary statements are taken as axioms. Any sentence which is a logical consequence of one or more of the axioms is also a sentence of that theory (theorem).
What are intuitive theories (Psychology)?

Intuitive theories are a special case of **scientific theories**, i.e. deductive theories that make falsifiable or testable predictions.

“**Systems of interrelated concepts that generate predictions** and explanations in particular domains of experience” (Murphy, 1993).

*Parent*: A person who has begotten or borne a child.  
*Child*: The offspring, male or female, of human parents.

Intuitive theories specify relations between objects

Imagine a set of identical-looking metal objects. Some pairs of object exert forces on each other when they come into close proximity.

“The causal laws are only defined in terms of the concepts, and the concepts are only defined in terms of the causal relationships between them.”
A problem of theory acquisition: Fodor's puzzle

**Standard view of concept learning**
New concepts are acquired by combining concepts that are already exist. (Laurence & Margolis, 2002).

Any given concept is either **unlearned** or **constructed** out of more primitive concepts.

**Fodor's puzzle:**
Most **lexical concepts** are unstructured and therefore unlearned primitives.

But the claim that concepts like *carburetor*, *coal*, and *electron* are not learned is highly counterintuitive.
Framework theories are particular intuitive theories

**Fundamental concepts** (chemicals, diseases, symptoms, animals)

**Possible relationships** between these concepts (chemicals cause diseases)
Binary relations can be represented as matrices

A relational system $R$ consisting of binary relations can be represented as a matrix.

$R : T^1 \times T^1 \rightarrow \{0, 1\}$

$T^1 = \{\text{heart disease, \ldots, bronchitis}\}$

“Specific theory”
(Relational system)
A framework theory organizes $\mathbb{R}$ into clean blocks.

Framework theories will not provide a complete explanation of any domain.

Their simplicity makes them a good initial target for models of theory learning.
Other binary relations with real-world applications

- Feedback loops
- Social relations
- Common cause reasoning
- Common effect reasoning
Multiple relations can be represented as cuboids

(a) Categories of people (unary relations \( T^1 \))
Relations between these categories (binary relations) \( R : T^1 \times T^1 \to \{0, 1\} \)

(b) Categories of people (unary relations \( T^1 \)), Categories of social predicates (unary relations \( T^2 \))
Relations between these categories (ternary relations) \( R : T^1 \times T^1 \times T^2 \to \{0, 1\} \)
Three main questions

- What are theories?
- How do they support inductive inference?
- How are they acquired?
How do theories support inductive inference?

Framework theories can be characterized as generative models that make inductive predictions about unobserved relations between entities.

**Bayesian inference** provides a principled framework for inference about category assignments.
A **theory** is an ordered pair \( T = (z, \eta) \), where \( z \) is a partition of entities into categories and \( \eta \) is a parameter matrix specifying the probability of relations.

A **relational system** \( R \) is a sample from a theory \( T \).

Entry \( R_{ij} \) is **generated by tossing a coin** with bias \( \eta_{AB} \), where \( A \) and \( B \) are the category assignments of entities \( i \) and \( j \).
Unobserved relations are predicted considering all $T$

A fully Bayesian learner considers the predictions of all possible theories.

$$P(R_{ij} = 1 | R) = \int P(R_{ij} = 1 | T) P(T | R) dT$$

Generative model for unobserved relations

Theory acquisition

**Method**
Approximate this integral by MCMC sampling of theories from their posterior distribution $P(T | R)$
Three main questions

- What are theories?
- How do they support inductive inference?
- How are they acquired?
How are theories acquired? – Previous approaches

Rational analysis of Anderson (1991):

Bayesian analysis of optimal estimates of unseen features under the assumption that features are independent within categories.

\[ P(R_{ij} = 1 | R) = \sum_{z_j} P(R_{ij} = 1 | z_j) P(z_j | R) \]

\[ z_j \quad ... \quad \text{category of person } j \]

people

features
Theories can be acquired using Bayesian inference

Given a formal characterization of a theory, a space of possible theories is set up and a prior distribution over this space is defined (favoring simple theories).

Bayesian inference then provides a normative strategy for selecting the theory in this space that is best supported by the available data.

\[ P(z, \eta|R) \propto P(R|\eta, z)P(\eta|z)P(z) \]
Prior distribution for partitions $z$

\[
P(z, \eta|R) \propto P(R|\eta, z)P(\eta|z)P(z)
\]

Chinese Restaurant Process (CRP)

\[
p(z_i = A|z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_A}{i-1+\gamma} & n_A > 0 \\ \frac{\gamma}{i-1+\gamma} & \text{A is a new category} \end{cases}
\]

$n_A$ … number of entities already assigned to category $A$

$\gamma$ … hyperparameter (number of categories)
Prior distribution for parameter matrices $\eta$

$$P(z, \eta|R) \propto P(R|\eta, z)P(\eta|z)P(z)$$

Beta function

$$P(\eta_{AB}|z) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \eta_{AB}^{\alpha-1} (1 - \eta_{AB})^{\beta-1}$$

$\Gamma(\cdot)$ … gamma function

$\alpha, \beta$ … hyperparameters (cleanness of blocks)
Likelihood function

\[ P(z, \eta|R) \propto P(R|\eta, z)P(\eta|z)P(z) \]

**Bernoulli distribution**

\[ P(R|\eta, z) = \prod_{ij} \eta_{c_i c_j}^{R_{ij}} (1-\eta_{c_i c_j})^{1-R_{ij}} \quad R_{ij} \in \{0, 1\} \]

\[ c_i, c_j \quad \text{... category (type) of item i and item j, respectively.} \]
Inference is carried out using MCMC methods

Inference is be carried out by sampling from the posterior on category assignments

\[ P(z|R) \propto P(R|z)P(z) \]

using

\[ P(R|z) \propto \int P(R|\eta z)P(\eta) d\eta \]

\[ P(R|z) = \prod_{A,B} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(m_{AB} + \alpha)\Gamma(\bar{m}_{AB} + \beta)}{\Gamma(m_{AB} + \bar{m}_{AB} + \alpha + \beta)} \]

where \( m_{AB} \) is the number of pairs \((i, j)\) where \( c_i = A, c_j = B \) and \( R_{ij} = 1 \), and \( \bar{m}_{AB} \) is the number of pairs where \( R_{ij} = 0 \).
The MAP partition is found using hill-climbing.

Operations move items from one category to another, split a category, or merge two categories to maximize the posterior on category assignments $P(z|R)$. 
Framework theories can also categorize features

\[ R^1 : T^1 \times T^2 \rightarrow \{0, 1\} \quad R^2 : T^1 \times T^1 \times T^3 \rightarrow \{0, 1\} \]

The approach can handle arbitrarily complex systems of features, entities and relations:

\[
P(z^1, z^2, \ldots, z^n, \eta^1, \ldots, \eta^m | R^1, \ldots, R^m) \\
\propto P(R^1, \ldots, R^m | \eta^1, \ldots, \eta^m, z^1, z^2, \ldots, z^n) \\
\times P(\eta^1, \ldots, \eta^m | z^1, z^2, \ldots, z^n) P(z^1, z^2, \ldots, z^n)
\]
Example: Animal categorization

O1 killer whale, blue whale, humpback, seal, walrus, dolphin
O2 moose, ox, sheep, buffalo, pig, cow
O3 antelope, horse, giraffe, zebra, deer
O4 hippo, elephant, rhino
O5 giant panda
O6 grizzly bear, polar bear
O7 german shepherd, tiger, leopard, fox, wolf, rat, weasel, bobcat, lion, raccoon
O8 skunk, mole, hamster, squirrel, rabbit, mouse
O9 dalmatian, persian cat, siamese cat, chihuahua, collie
O10 beaver, otter
O11 monkey, gorilla, chimp
O12 bat

F1 flippers, strain teeth, swims, arctic, coastal, ocean, water
F2 hooves, long neck, horns
F3 hands, bipedal, jungle, tree
F4 bulbous body shape, slow, inactive
F5 meat teeth, eats meat, hunter, fierce
F6 walks, quadrupedal, ground
F7 orange, red, yellow, stripes, flies, hops, tunnels, eats insects, scavenger, desert, cave
F8 pads, claws, nocturnal, hibernate, stalker
F9 black, brown, furry, chew teeth, new world
F10 blue, tusks, eats plankton, skimmer
F11 white, patches, spots, domestic
F12 fast, active, agility
F13 forager, forest, nest spot
F14 tail, old world
F15 plains, fields
F16 big, strong
F17 paws, solitary
F18 bush, mountains
F19 lean
F20 grazer
F21 eats fish
F22 muscle
F23 tough skin
F24 smart
F25 hairless
F26 smelly
F27 timid
F28 vegetation
F29 buck teeth
F30 gray
F31 weak
F32 small
F33 long leg
F34 group
Relationships between animal and feature types

**Strong relationships**

Aquatic mammals
- “has flippers”, “swims”, “lives in the ocean”

Primates
- “has hands”, “lives in the jungle”, “found in trees”
Feature-based models lack relations of features between categories

Feature-based models assume the features are conditionally independent given the set of item categories (people).

Learning that a robin has wings and flies provides evidence that other robins are likely to fly, but does not support the conclusion that winged entities from other categories are likely to fly.
Most of the categories of the feature-based model can be created by merging two or more of the categories discovered by the relation based model.
Every relational system $R$ can be converted in a feature-based model

\[ R^1: T^1 \times T^1 \times T^2 \rightarrow \{0, 1\} \quad \Rightarrow \quad \]

\[ R^2: T^1 \times T^3 \rightarrow \{0, 1\} \]

Features in $T^3$ are all items:

\[ R^1(a, x^1, x^2) \]
\[ R^1(x^1, a, x^2) \quad x^1 \in T^1, \quad x^2 \in T^2 \]
A model that discovers simple theories or systems of related concepts was presented.

The model simultaneously discovers the concepts that exist in a domain and the laws or principles that capture relationships between these concepts.

It was demonstrated that statistical inference can help to explain the acquisition of highly-structured representations: as sophisticated as intuitive theories.

... but how well does it work?