Learning overhypotheses with hierarchical Bayesian models

Charles Kemp Amy Perfors Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, USA

Presented Tuesday, April 15, 2008

Goodman and His Magic (Bayesian?) Marbles

(Losing?) Goodman's Marbles

- Goodman introduces overhypotheses with an example based on bags of colored marbles (cite Goodman 1955)
- Suppose we have a stack of bags filled with colored marbles
- Empty several bags and find some contain all black marbles and the rest all white marbles
- Choose a new bag and draw a single black marble

Hypotheses and Overhypotheses

- How many black marbles do you think are in the bag?
  - *H*: All of the marbles
  - *H* is an example of a *hypothesis*
- Why do you think this?
  - *O*: Each bag in the stack contains marbles that are all the same in color
  - *O* is an example of an *overhypothesis*

Hypotheses and Overhypotheses

- How many black marbles do you think are in the bag?
  - *H*: All of the marbles
  - *H* is an example of a *hypothesis*
- Why do you think this?
  - *O*: Each bag in the stack contains marbles that are all the same in color
  - *O* is an example of an *overhypothesis*

Hypotheses and Overhypotheses

- How many black marbles do you think are in the bag?
  - *H*: All of the marbles
  - *H* is an example of a *hypothesis*
- Why do you think this?
  - *O*: Each bag in the stack contains marbles that are all the same in color
  - *O* is an example of an *overhypothesis*

Hypotheses and Overhypotheses

- How many black marbles do you think are in the bag?
  - *H*: All of the marbles
  - *H* is an example of a *hypothesis*
- Why do you think this?
  - *O*: Each bag in the stack contains marbles that are all the same in color
  - *O* is an example of an *overhypothesis*

Defining Overhypotheses

- For our purposes an overhypothesis is 'any abstract knowledge that sets up a hypothesis space at a less abstract level'
- *O* is an overhypothesis since it sets up a space of hypotheses about the marbles in the bag
- It say the marbles can be
  - all black,
  - all white,
  - all green,
  - etc.

Marble Problem

- Marble problem suggests the two-fold problem of inference
- A theory of learning must answer
  - How do people generalize from data to hypotheses, e.g. from seeing a few marbles to a hypothesis about the distribution of the bag
  - How do people learn an overhypotheses which allows us to generalize this knowledge which aid in the generalization task, e.g. the "uniform color" hypothesis

Hierarchical Bayesian Models

- Hierarchical Bayesian models (HBMs) offer one potential way for dealing with both of these questions simultaneously
- These models can capture data generated in a step-wise fashion by positing each of these steps as a level in a hierarchy
- For example the overhypothesis from the marble problem sets up a two-step generating process
  1. Pick a color
  2. Fill a bag entirely with that color marble

Bayes' Rule

- Bayes' rule is a statistical method for relating knowledge regarding parameter values to observed data
- It does this by a forming posterior distribution, $\pi(\theta|X_1, \ldots, X_n)$, given
  - a prior distribution, $\pi(\theta)$, and
  - a likelihood, $p(X_1, \ldots, X_n|\theta)$
- It states

$$\pi(\theta|X_1, \ldots, X_n) = p(X_1, \ldots, X_n|\theta)\pi(\theta)$$

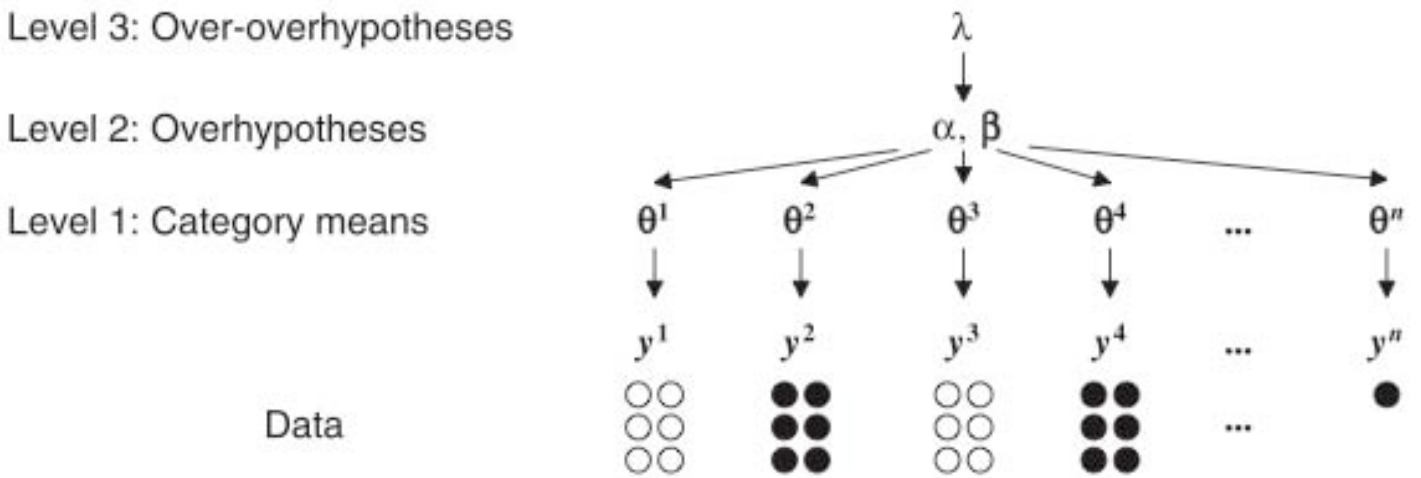
From Bayes' Rule to HBMs

- Bayes' rule can be made hierarchical by placing a prior distribution over potential prior distributions
- In the marble problem this means that corresponds to the placing a prior distribution on potential overhypotheses

Returning (to) Goodman's Marbles

- What sorts of information should we use to characterize our overhypothesis about the bags of marbles?
  - The extent to which bags are uniform
  - Distribution of colors across the entire collection of bags
- There may be other ways of characterizing overhypotheses about the bags; however, one of the strengths of Bayesian inference is its ability to handle different assumptions about the structure of the overhypotheses

Slightly More Formally...

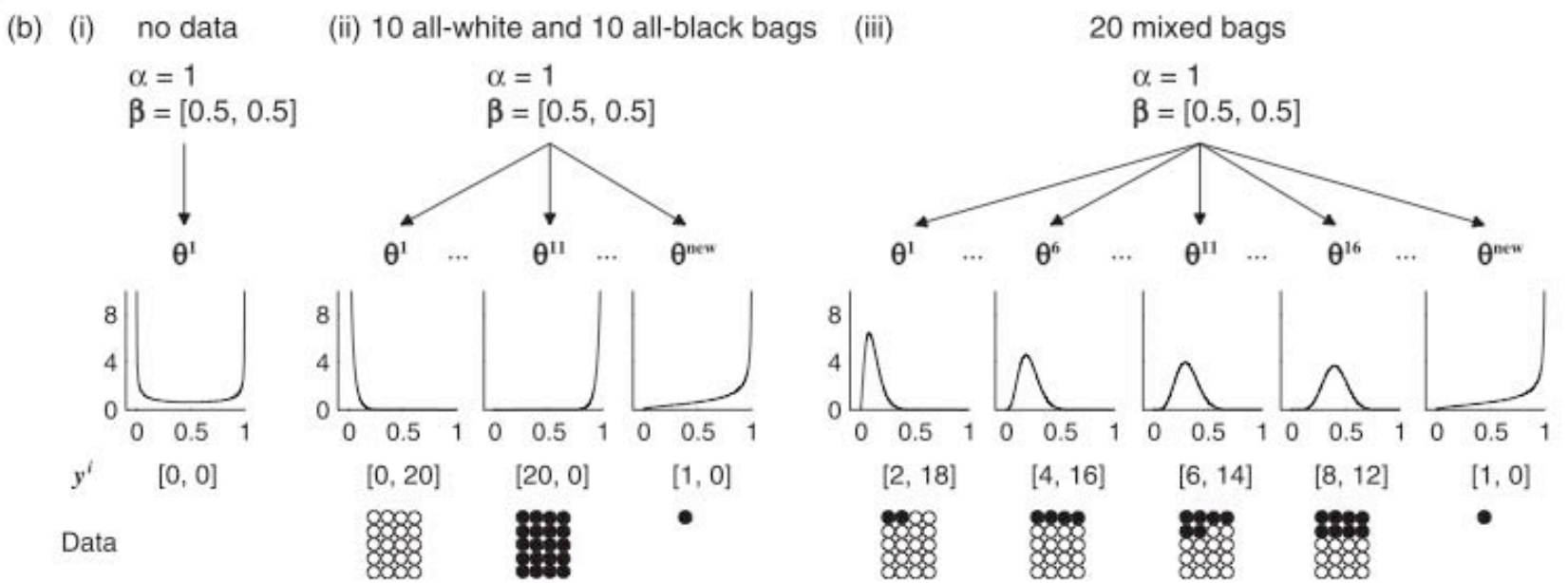


- $\boldsymbol{y}^{\boldsymbol{i}}$: set of observations of marbles in bag *i*
- $\boldsymbol{\theta}^{\boldsymbol{i}}$: true color distribution for *i*th bag in stack
- α: extent to which colors in bag are uniform
- β: distribution of colors across entire collection of bags
- λ: 'innate knowledge'?

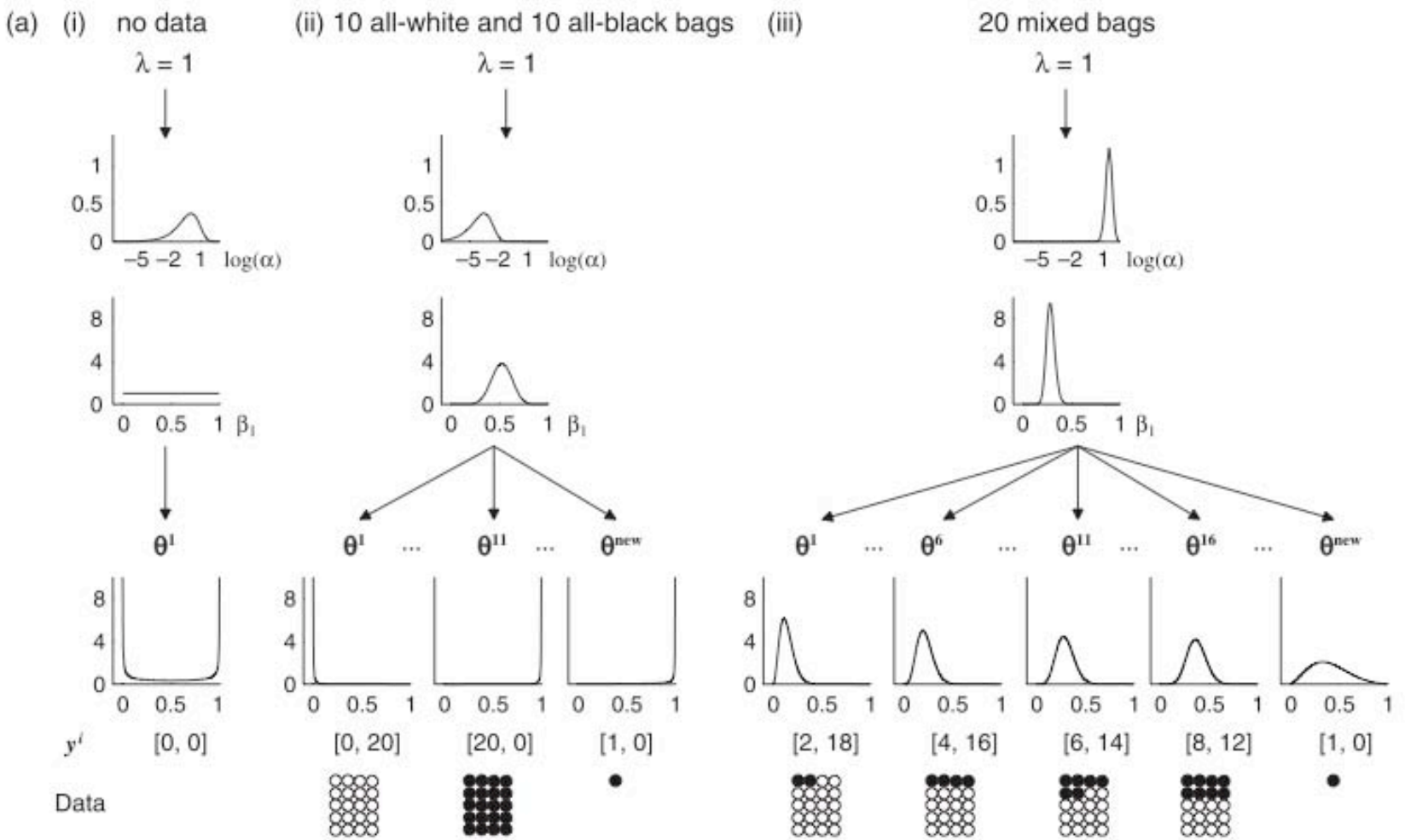
λ as Innate?

- The article glosses over λ, the 'over-overhypothesis', saying it might be 'innate knowledge'
- The important thing to note about λ is that the data washes out its effect on inference, making it not matter (except in the way it parameterizes more basic levels)
- In terms of innateness λ has more in common with the process by which axons grow or mylinate than with a innate grammatical construction

Two Steps Forward... (Generation)



...One Step Back (Inference)



I'll Show YOU a 'Shape Bias'

1^{st} and 2^{nd} Order Generalization

- 1^{st} order generalization:
  - Present a novel exemplar from a previously learned category
  - Ask participant to choose another object in same category
- 2^{nd} order generalization:
  - Present a novel exemplar from a novel category
  - Ask participant to choose another object in same category

What's (in) a 'Shape Bias'

- The 'shape bias' is a bias toward making these generalization on the basis of an exemplar's shape
- Given a single exemplar of a novel object category, children extend the category label to similarly shaped objects ahead of objects that share the same texture or color (cites)

From Colored Marbles to Shaped 'Zups'

- The key to this article is knowing how uniformly colored marbles could possible map onto the shape bias
- 'Color' is a *feature* of the marble, roughly a subordinate property that marbles can have
- Likewise shape, size, texture, etc. are features that 'zups' can have
- Just like we defined an overhypothesis about marble color we can define an overhypothesis about shape

The Shape Overhypothesis

- In words we might give a 'shape overhypothesis': shape tends to be homogenous within object categories
- In words this sounds a lot like our marble color overhypothesis: color tends to be homogenous within bags of marbles
- This suggests we might be able to model it using a similar model

Feature Generalization

- The shape bias is part of two related larger psychological questions:
  - How do people generalize feature possession from one member of a category to another, e.g. within dogs, if you know a german shepard has a given feature, how do we infer whether a pug has it?
  - How do people generalize novel objects to categories, e.g. if we know an animal has four legs and barks, how do we infer it's a dog?

Generalization and Overhypotheses

- The larger questions on generalization suggest two types of overhypotheses may be useful to human learners
  - How homogenous a given feature is within a given category
  - How prevalent is a feature among the population at large
- These are precisely the aspects of a feature we modeled in the marble example
  - Within category homogeneity was α
  - Prevalence in the population was β

The Shape Bias Demystified?

- This HBM of feature generalization has the shape bias as a consequence
- 'Shape tends to be homogenous within object categories' maps onto low values of α
- This overhypothesis makes it unlikely that an exemplar with a different shape than *A* comes from the same category as *A*

Results

(a) Training

Category	1 1	2 2	3 3	4 4
Shape	1 1	2 2	3 3	4 4
Texture	1 2	3 4	5 6	7 8
Color	1 2	3 4	5 6	7 8
Size	1 2	1 2	1 2	1 2

(b) First-order generalization

T_1

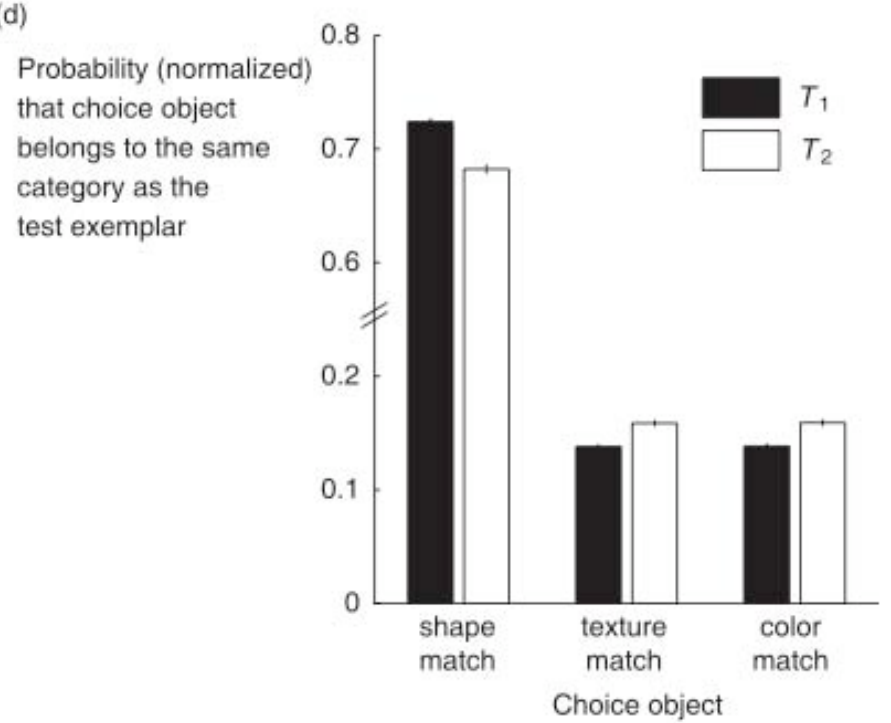
1	?	?	?
1	1	6	6
1	9	1	9
1	9	9	1
1	1	1	1

(c) Second-order generalization

T_2

5	?	?	?
5	5	6	6
9	10	9	10
9	10	10	9
1	1	1	1

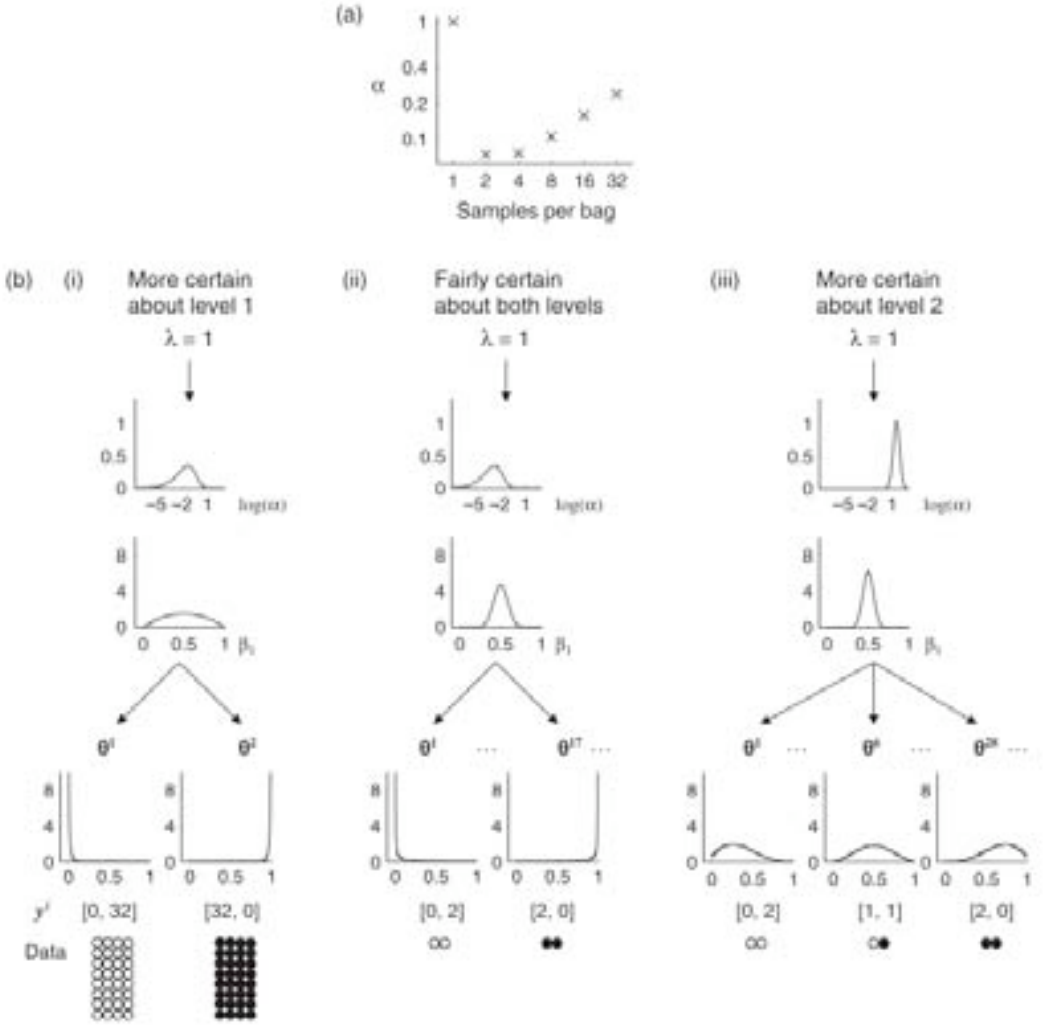
(d)



HBMs and Parallel Learning

- HBMs have the capacity to learn about multiple levels in the hierarchy in parallel
- This leads to at least three possible outcomes
  - More certain about Level 1 (hypotheses) than Level 2 (overhypotheses) knowledge
  - Relatively confident about both levels
  - More certain about Level 2 than Level 1 knowledge

Learning Possibilities



Wealth in Poverty?

- This example suggests that given certain types of data less data may be better than more
- As one of your questions noted, this suggests that the 'poverty of the stimulus' might actually be a good thing if the small amount of data we get is good data
- Roughly this suggests that consistent results across a large number of categories is better for learning overhypotheses than results that are only consistent across a small number
- Are linguistic stimuli like this?

An Ontological Kind of Chinese Restaurant Process

Ontological Kinds

- *Ontology* (from Dictionary.com):

  A rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations

- Might call *ontological kinds* groups of exemplars which share similar structural properties

Learning Ontological Kinds

- The shape bias discuss suggests how children might learn 'ontological kinds'
- If you know a set of feature across which exemplars vary there exist statistical models which are able to segregate exemplars based on different degrees of homogeneity and prevalence
- One such model is called a 'Chinese restaurant process'

Chinese Restaurant Processes

- Describing the mathematical problem that Chinese restaurant processes solve gets quite technical; however, understanding on some level what they do is important
- Imagine a Chinese restaurant with an infinite number of circular tables with infinite capacity
- Every minute a new customer walks in and decides at random which of the tables to sit at
- Uniformly at random he or she can decide to:
  - Sit directly to the left of one of the customers already there
  - Sit at a new, unoccupied tables
- The goal of inference is to get a better idea where customers should actually sit

Exemplars Get Dim Sum

- The Chinese restaurant process suggests a 'rational' framework for learning ontological kinds
  - We see a few exemplars, these are customers in the restaurant
  - The tables are ontological kinds
  - *A priori* any exemplar can sit in any ontological kind
  - Inference determines the probabilities over which exemplars should sit together and which shouldn't
- This type of inference does not require kinds to be defined *a priori*, the properties of the exemplars will 'determine' the kinds

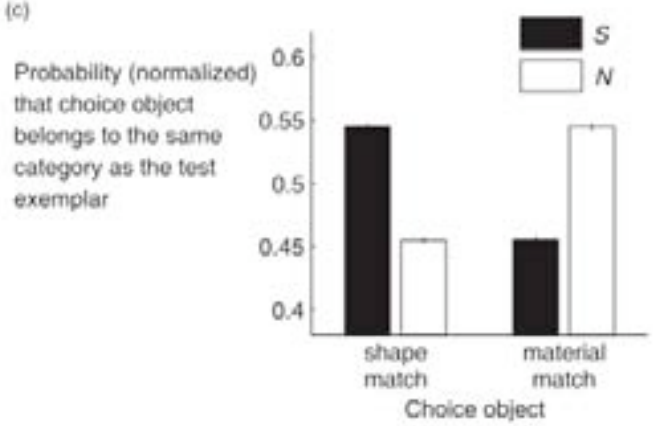
Results of Learning

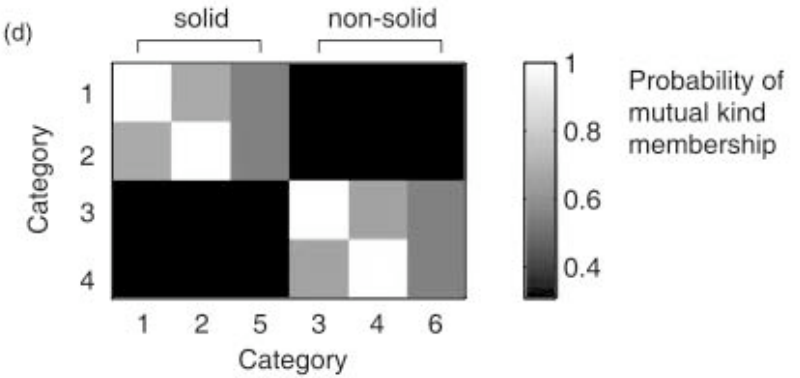
(a) Training

Category	1	1	2	2	3	3	4	4
Shape	1	1	2	2	3	4	5	6
Material	1	2	3	4	5	5	6	6
Size	1	2	1	2	1	2	1	2
Solidity	1	1	1	1	2	2	2	2

(b) Second-order generalization

S			N		
5	?	?	6	?	?
7	7	8	8	8	9
7	8	7	8	9	8
1	1	1	1	1	1
1	1	1	2	2	2





Learning Grammars

- This model suggests ways which grammar and syntax may be learned
- Given an appropriate parameterized model, language learners parse overhypotheses based on linguistic stimuli
- These overhypotheses form into the rules of the grammar

Problems for HBMs

- Can grammars be explained solely in terms of parameterized models?
- The current account requires the features along which exemplars vary to be known *a priori* but these features are themselves a source of uncertainty
- The current account parameterizes particular aspects of features but this parameterization is also potentially a source of uncertainty for the learner

Questions From the Peanut Gallery

(Q) The over-overhypothesis doesn't seem to play a very important role in the models that are discussed in this paper, but I'm wondering if there are other models in which different values for the over-overhypothesis makes more of a difference?

(Q) There is an example in the paper schematized in figure 1a.) regarding the three levels of knowledge, and how they are related. The paper mentions how there could be theoretically n levels of overhypothesis' which each guide inference on the next 'more basic' level down. But, what is to suppose, then, that the most basic level of inference is not dependant on more than just the next highest level, where examples and evidence can be observed. I don't see any reason why, say, the i-th level cannot be a function of the (i+1, ... n)th levels, especially since doing so may provide more 'accurate' generalizations and inferences besides theoretical simplicity.

(Q) The description of HBM shows that children are able to infer category and hierarchy from concrete tasks involving visual perception and discrete objects (marbles, tribesmen etc.) and ascriptive characteristics. How would the authors apply Hierarchical Bayesian modeling to the problems of parsing and of the grammatical correctness of novel syntax? It seems like there needs to be concrete conditions or steady states that can be analyzed or appraised by an HBM model Ð does language provide sufficient conditions for such a model to take place?

(Q) The paper argues that overhypothesis of some level are innate, but I struggle to see how this can necessarily apply to children, rather than it being an algorithmic, and essentially academic, discussion. Suppose we do find a learning algorithm which models precisely at some degree the acquisition of a child's language with some set of innate knowledge (the algorithms, and other things), is this still enough to claim that it is how it actually works? Certainly this kind of work has applications in artificial intelligence-related endeavors, because it bloats (for lack of a better word) the kind of inductive reasoning that we might usually use. But I guess I still am a bit unsettled with the idea of proposing an algorithm and calling it a 'child learning' hypothesis.