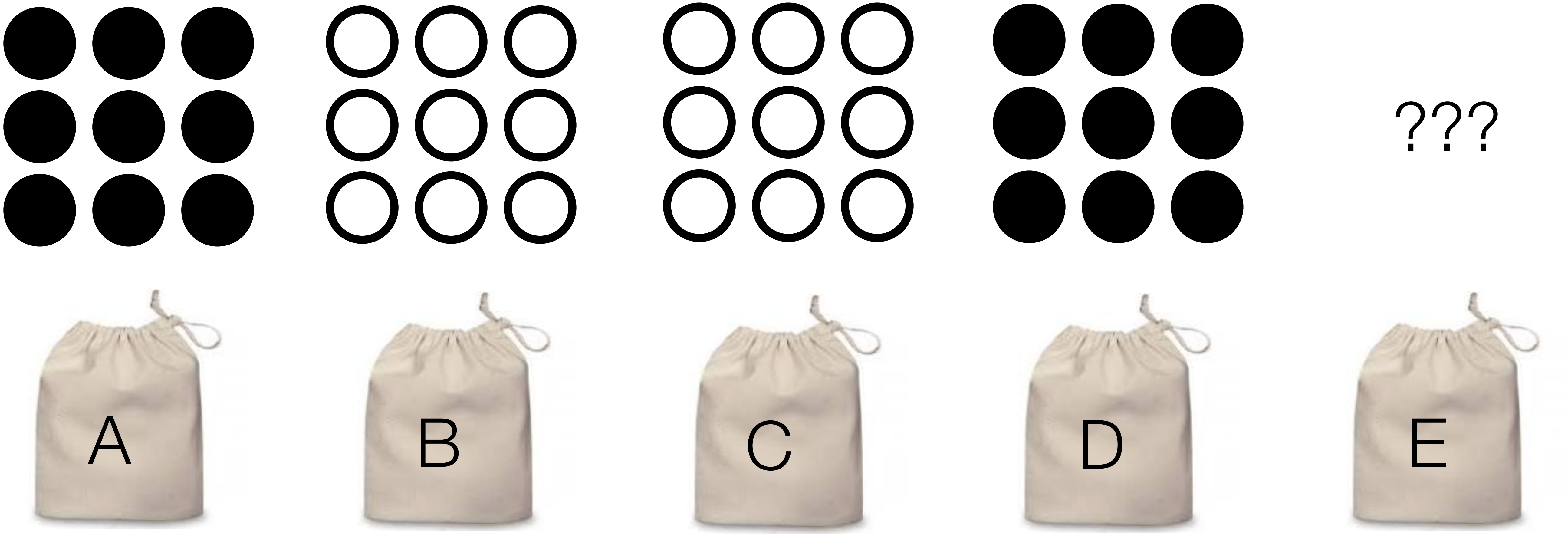
Simulating Language
7: Hierarchical models and learning to learn

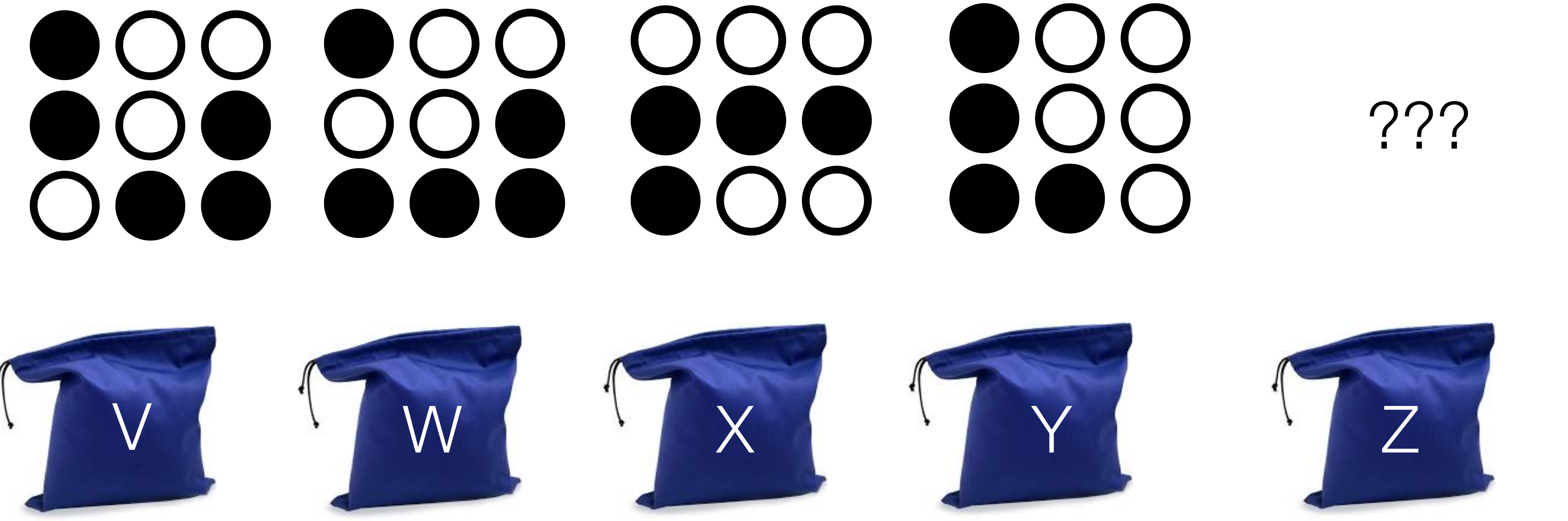
Kenny Smith
kenny.smith@ed.ac.uk



Option 1: Bag E contains marbles, beyond that I cannot say

Option 2: Bag E contains a mix of roughly equal numbers of black and white marbles

Option 3: Bag E contains either exclusively black marbles, or exclusively white marbles



Option 1: Bag Z contains marbles, beyond that I cannot say

Option 2: Bag Z contains a mix of roughly equal numbers of black and white marbles

Option 3: Bag Z contains either exclusively black marbles, or exclusively white marbles

The prior

$$P(h \mid d) \propto P(d \mid h)P(h)$$

Priors include

- Expectations about word meanings (week 3)
- Expectations about regularity / variability (weeks 4-5)
- Expectations about degeneracy / holism / compositionality (week 7)

Where does the prior come from?

$$P(h|d) \propto P(d|h)P(h)$$

- Could be due to very general constraints on learning (e.g. the simplicity prior used last week)

- Could be due to learning in another domain (e.g. a regularity preference because you’ve learned the universe tends to be predictable?)

- Could be domain-specific expectations that you are somehow born with (see upcoming weeks for a model of this!)

- **Could be learned domain-specific expectations**

Motivating examples involving language, not marbles: reminder of some stuff from lecture 2

Quine (1960): meaning underdetermined by data

- The four legged animal
- The two legged animal
- Some part of either (the leg, the hat, …)
- Some property of some part (the length of the leg, the material of the hat)
- Nothing to do with what you're seeing ("I'm hungry")
- Something weirder (a wet nose and a waggable tail, but only until Scotland win the World Cup)

There are in principle **infinitely many possible meanings** for "doggy" which would be consistent with this usage, and **any possible sequences of usages**

Learners must have **some** constraints on word meaning

Minimally: to rule out the extremely wacky word meanings

But maybe they are more detailed:

- Expectations about meanings (e.g. words refer to whole objects, words refer to basic-level categories, words generalise by shape of referent, …: Macnamara, 1972; Markman, 1989; Landau, Smith & Jones, 1988)

- Expectations about words (e.g. word meanings are mutually exclusive: Markman & Wachtel, 1988)

- …

Learners must have **some** constraints on word meaning

Minimally: to rule out the extremely wacky word meanings

But maybe they are more detailed:

- Expectations about meanings (e.g. words refer to whole objects, **words refer to basic-level categories**, words generalise by shape of referent, ….: Macnamara, 1972; Markman, 1989; Landau, Smith & Jones, 1988)
- Expectations about words (e.g. word meanings are mutually exclusive: Markman & Wachtel, 1988)
- …

Learners must have **some** constraints on word meaning

Minimally: to rule out the extremely wacky word meanings

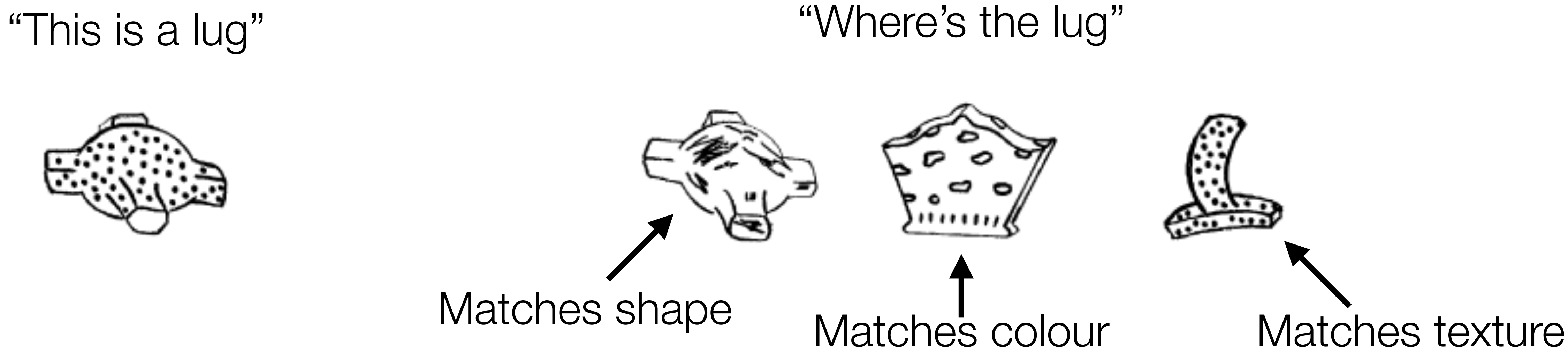
But maybe they are more detailed:

- Expectations about meanings (e.g. words refer to whole objects, words refer to basic-level categories, **words generalise by shape of referent**, …: Macnamara, 1972; Markman, 1989; Landau, Smith & Jones, 1988)
- Expectations about words (e.g. word meanings are mutually exclusive: Markman & Wachtel, 1988)
- …

The shape bias

- In English, shape of objects is the most reliable cue to category membership, and therefore the most reliable cue to object names
    - i.e. concrete count nouns tend to generalise by shape, not texture, colour, material etc: cups are cup-shaped, chairs are chair-shaped, trousers are trouser-shaped, …

- Children aged 3+ seem to be aware of this, and systematically generalise new object names by shape (e.g. Landau et al., 1988): **the shape bias**



“This is a lug”

“Where’s the lug”

Matches shape

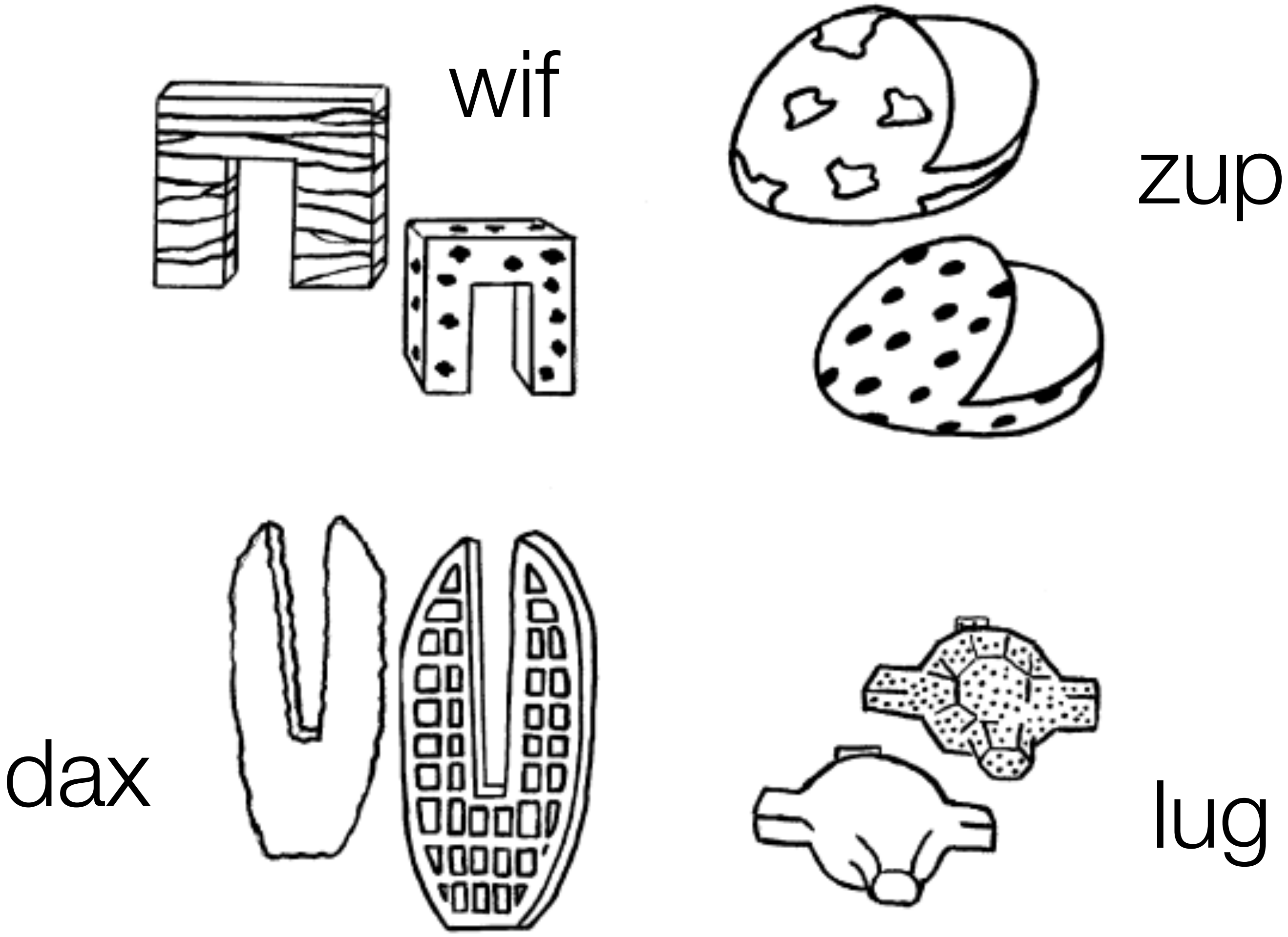
Matches colour

Matches texture

Learning the shape bias (L. Smith et al, 2002)

- 18 month old English-speaking children (i.e. too young to show the shape bias)

- Experimental group get 7 week training programme on novel objects whose labels generalise by shape

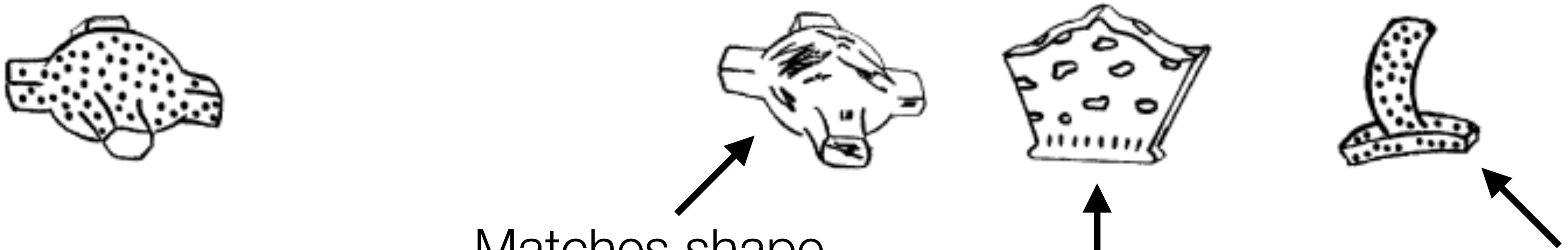


Learning the shape bias (L. Smith et al, 2002)

- Week 8: first-order generalisation test with trained label and 3 novel objects

"This is a lug"

"Where's the lug"



Matches shape

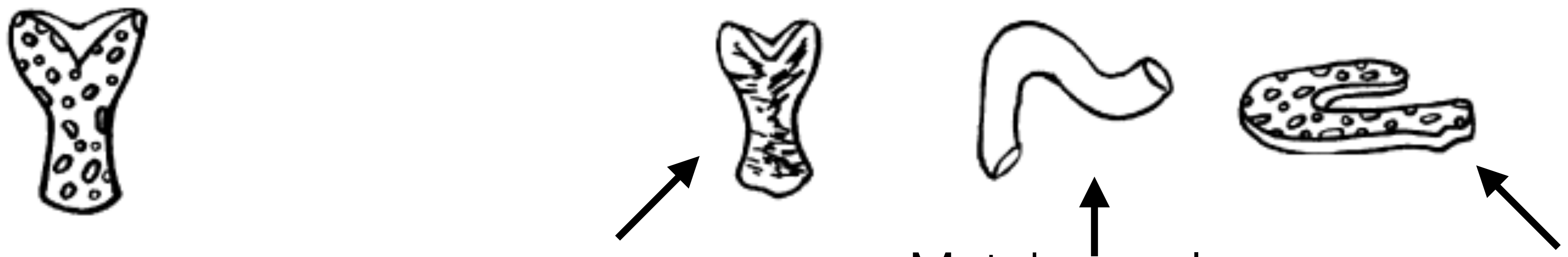
Matches colour

Matches texture

- Control group: 36% generalise by shape (i.e. chance)
- Trained children: 88% generalise by shape

Learning the shape bias (L. Smith et al, 2002)

- Week 9: second-order generalisation test with **novel** label and 3 novel objects



“This is a veet”

“Where’s the veet”

Matches shape

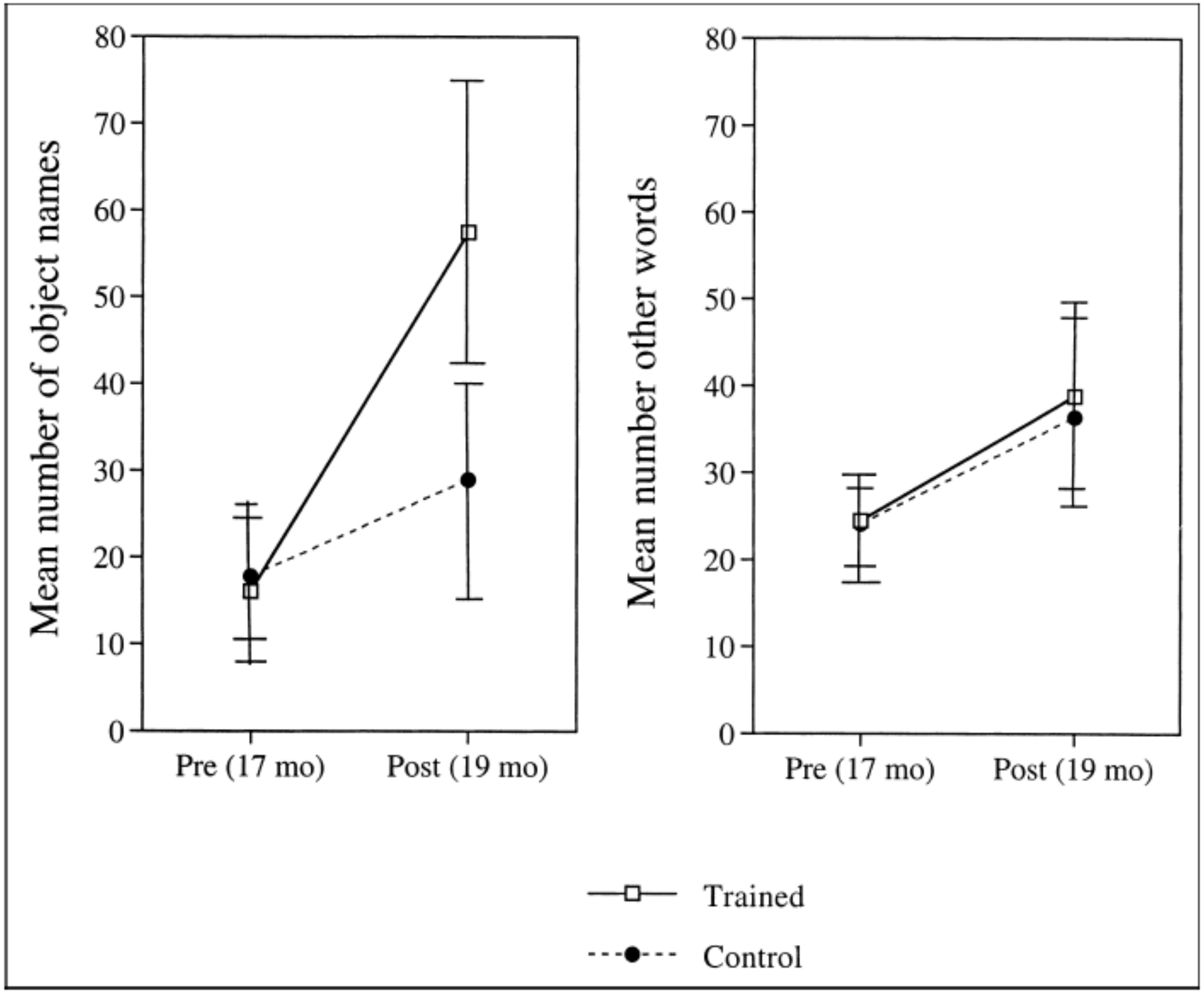
Matches colour

Matches texture

- Control group: 34% generalise by shape (i.e. chance)

- Trained children: **70% generalise by shape**

Learning the shape bias (L. Smith et al, 2002)



How do we capture this in a model?

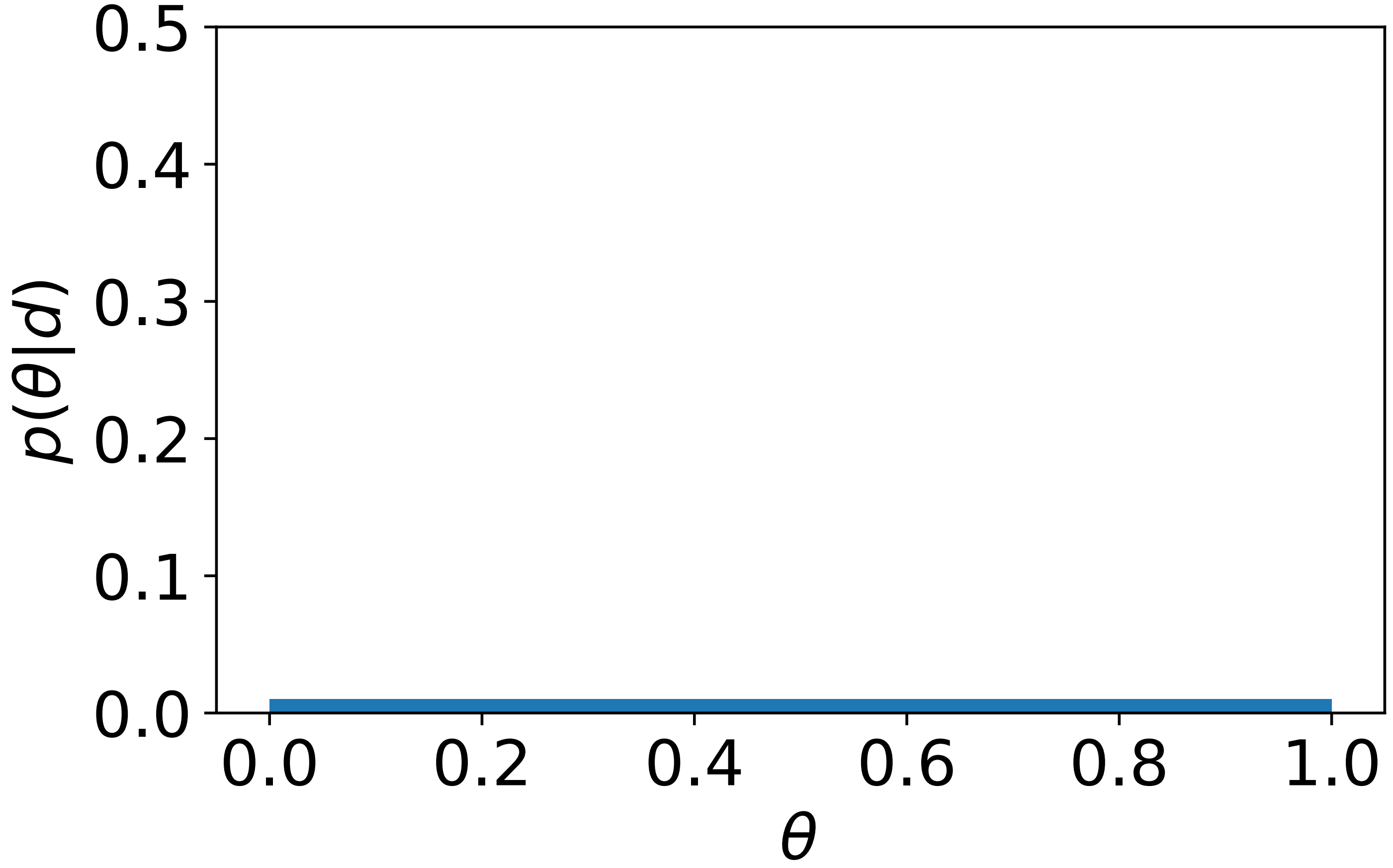
- Rather than being fixed, the prior is itself learned (and the learned prior can therefore guide subsequent learning)

- We can model learning the prior as a process of Bayesian inference in the usual way

- Of course this means we need a prior over our prior, which is why these models are called **hierarchical**

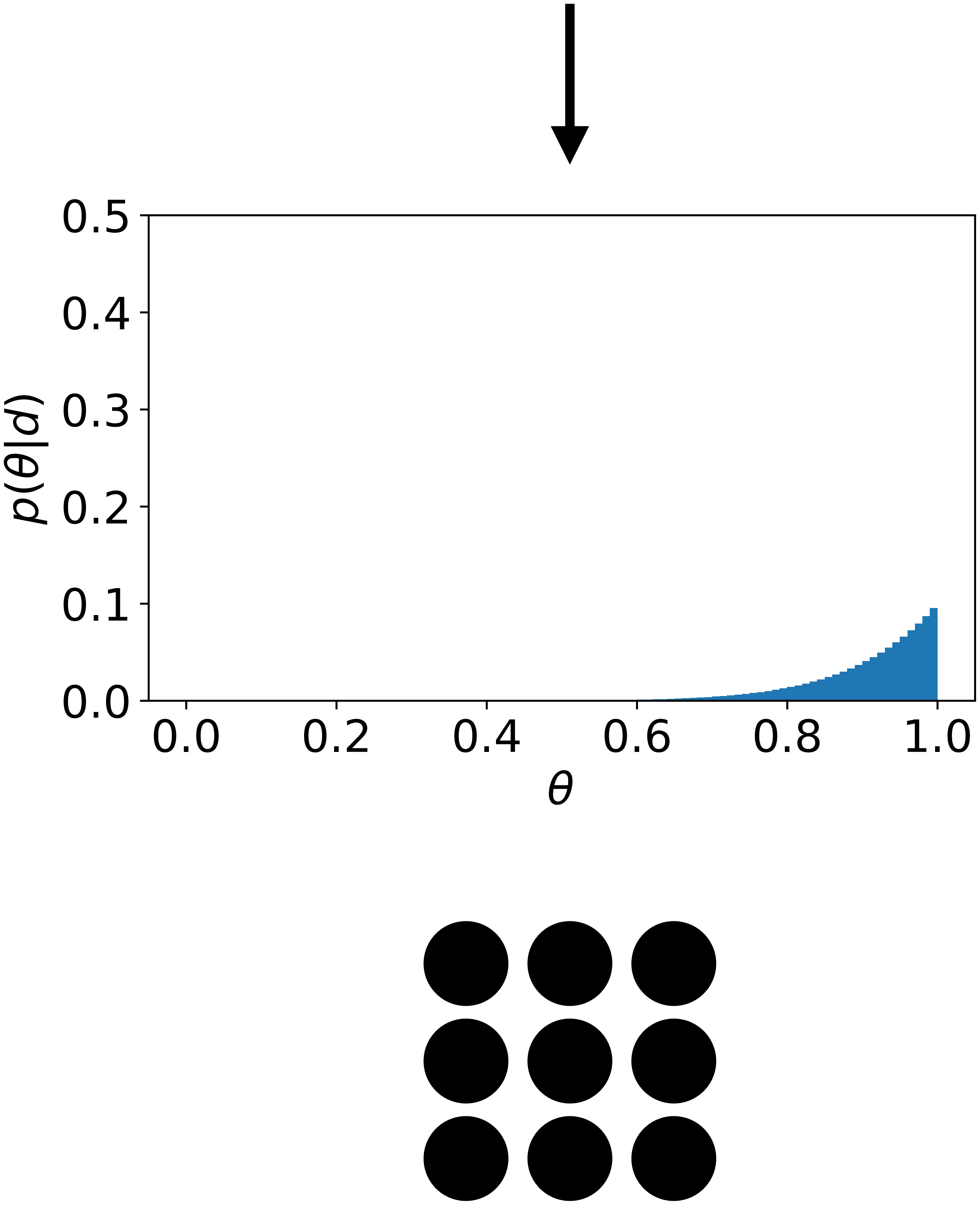
Level 3	γ
Level 2	α
Level 1	θ
Data	d

$\alpha = 1$

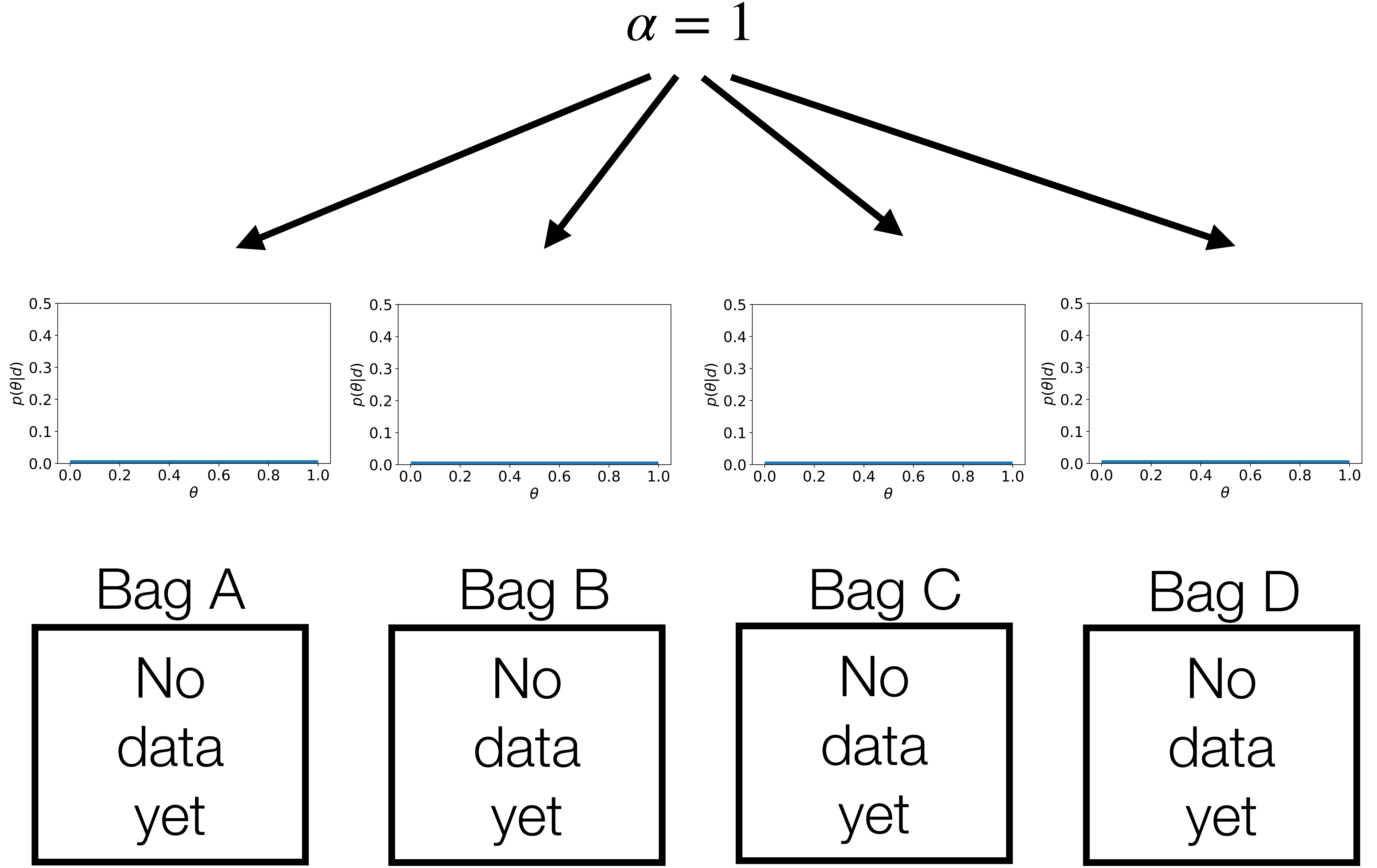


No
data
yet

$\alpha = 1$



$\alpha = 1$



Bag A

No data yet

Bag B

No data yet

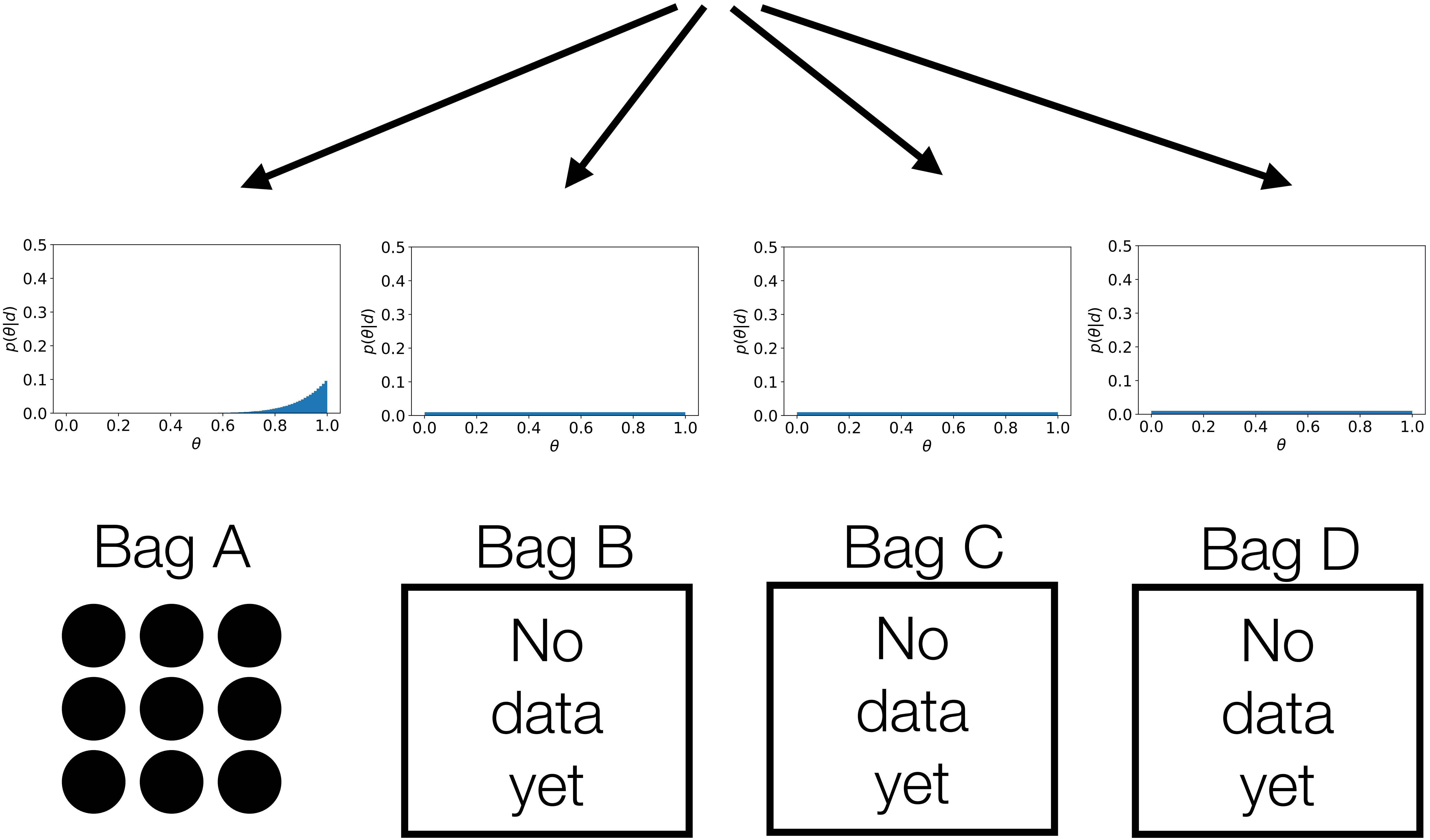
Bag C

No data yet

Bag D

No data yet

$\alpha = 1$



Bag A

Bag B

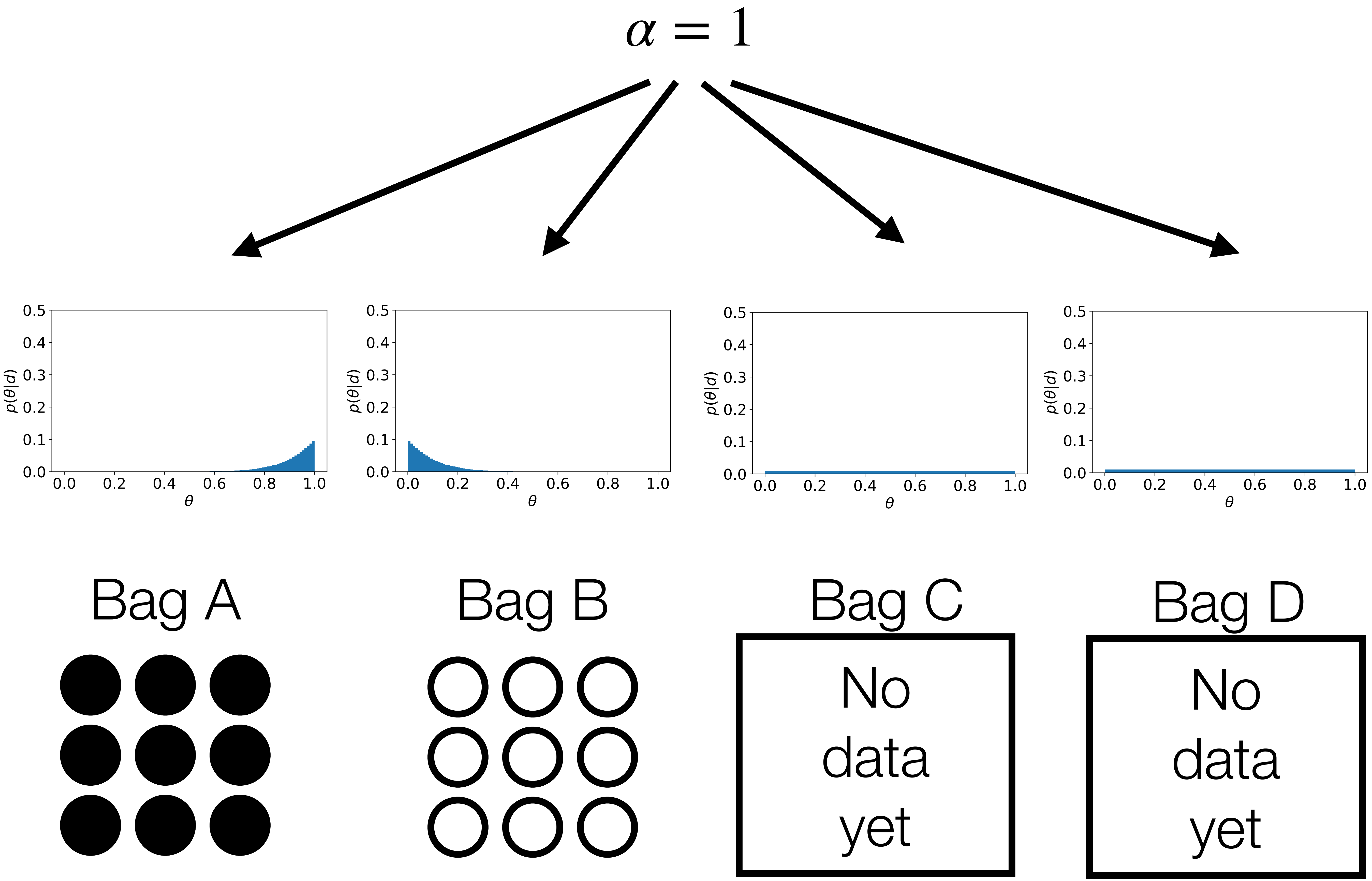
No data yet

Bag C

No data yet

Bag D

No data yet



$\alpha = 1$

Bag A

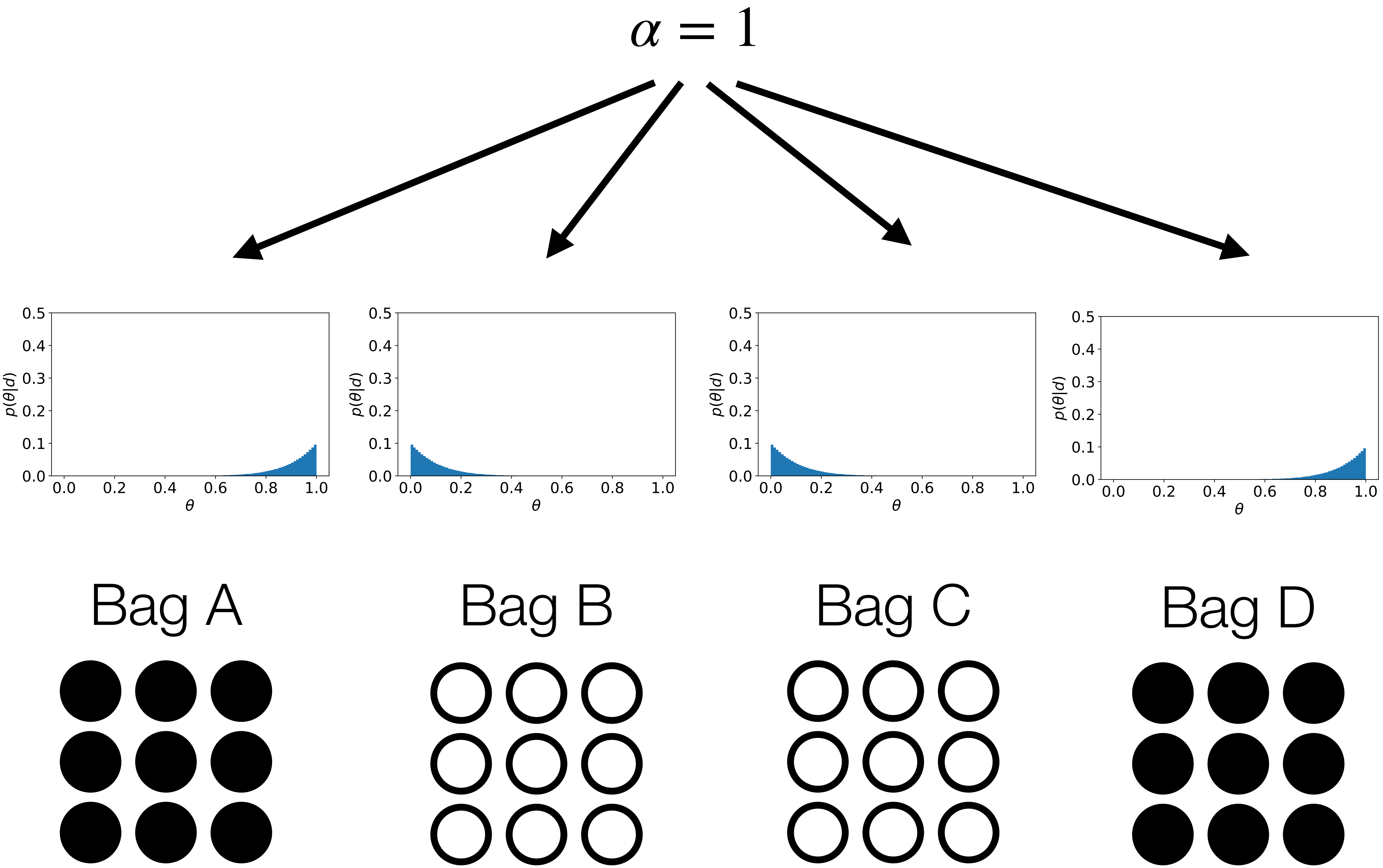
Bag B

Bag C

No data yet

Bag D

No data yet



$\alpha = 1$

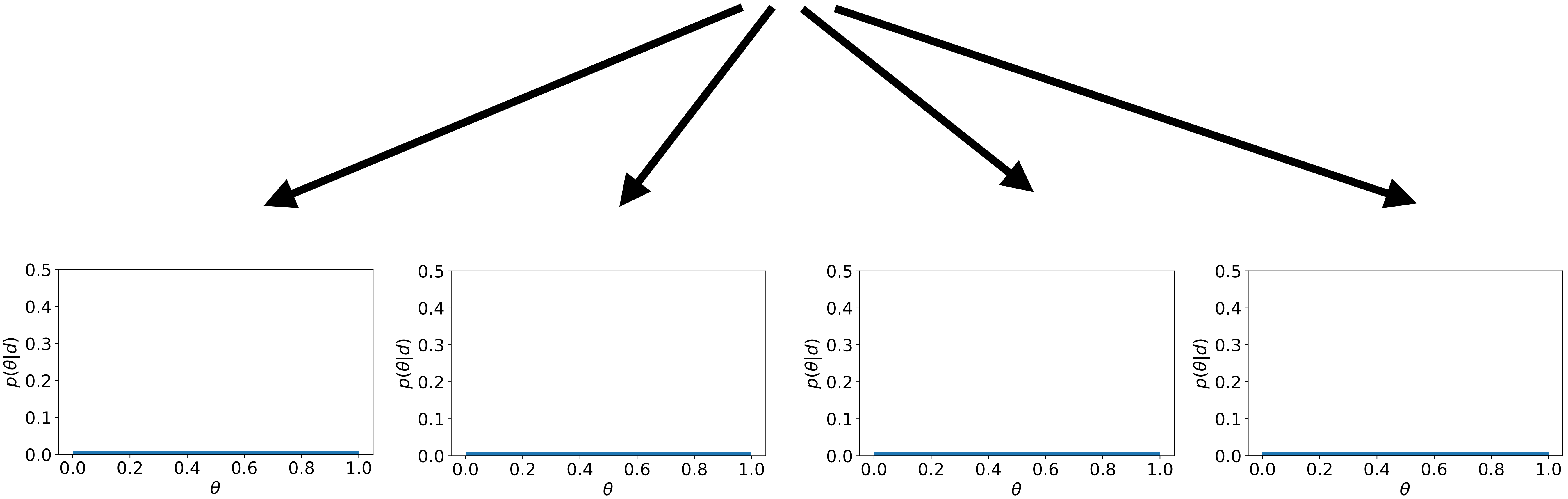
Bag A

Bag B

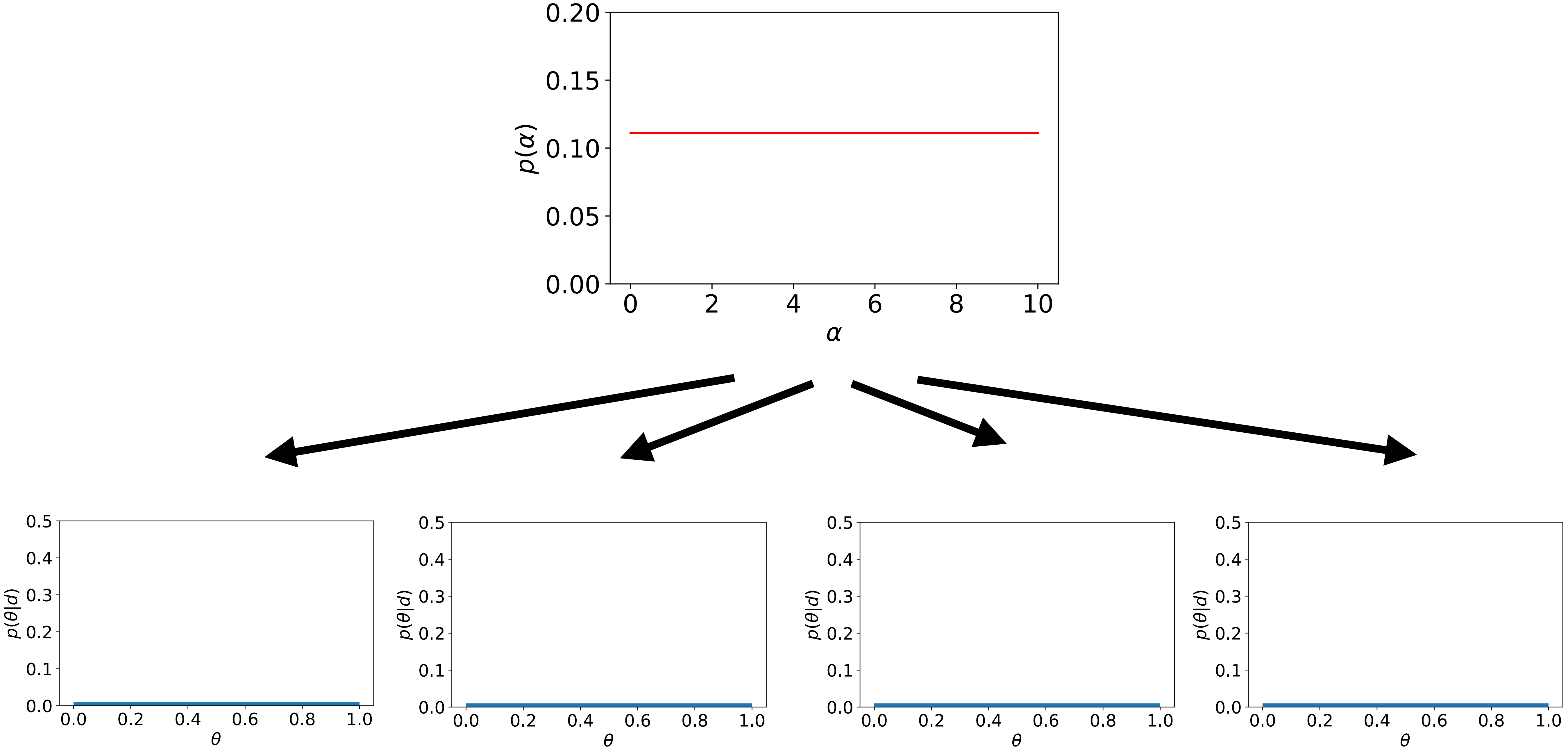
Bag C

Bag D

$\alpha = ???$



Bag A	Bag B	Bag C	Bag D
No data yet	No data yet	No data yet	No data yet



Bag A

No data yet

Bag B

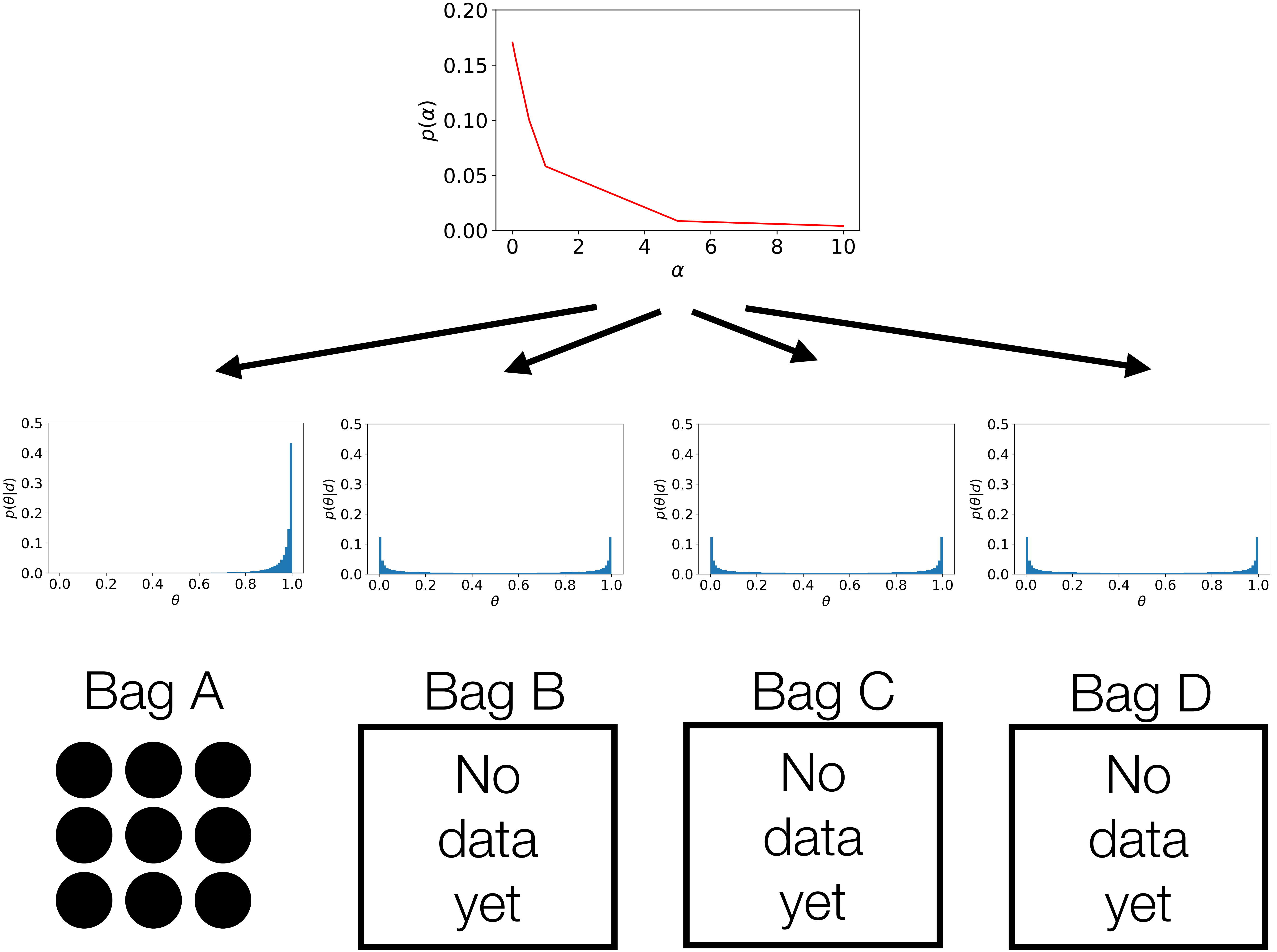
No data yet

Bag C

No data yet

Bag D

No data yet



Bag A

Bag B

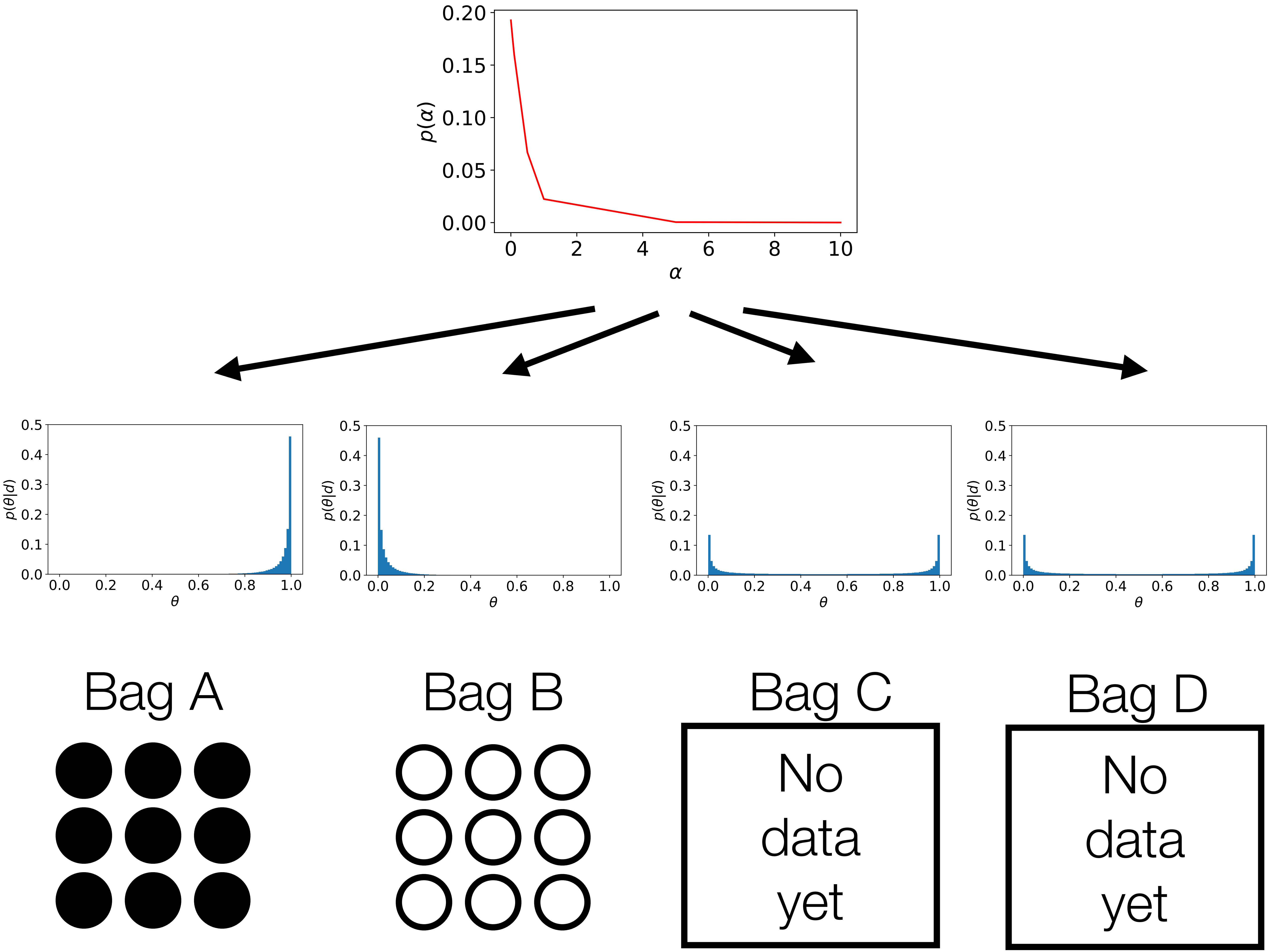
No data yet

Bag C

No data yet

Bag D

No data yet



Bag A

Bag B

Bag C

No data yet

Bag D

No data yet

The same thing in maths for those that prefer it

The familiar non-hierarchical model

$$P(\theta | d) \propto P(d | \theta) P(\theta | \alpha)$$

Hierarchical model, inferring α

$$P(\alpha | d) \propto \int_{\theta} P(d | \theta) P(\theta | \alpha) P(\alpha)$$

Hierarchical model, inferring θ

$$P(\theta | d) \propto \int_{\alpha} P(d | \theta) P(\theta | \alpha) P(\alpha)$$

These learned biases are probably **everywhere**

Just a hunch, but I think we might be massively underestimating the power of learned biases to shape learning and explain the surprising precocity of language learners

- Basic level bias, shape bias, …
- Mutual exclusivity - develops over time (Halberda, 2003), is weaker in bilingual children (Houston-Price et al., 2010)
- Syntactic categories
- Correlations between semantic/phonological cues and syntactic category (e.g. in English, nouns tend to be longer than verbs, 4-year-olds know this: Cassidy & Kelly, 1991)
- Pragmatic inference?
- Structure dependence in syntax??
- …

Summary and next up

- Priors can be learned
- We can capture this as Bayesian inference, using a hierarchical model
- There is strong evidence that humans learn to learn in this way
- Several options available on the readings page for this lecture, from brief and non-technical to long and somewhat technical
- Thursday and Friday: lab on a simple hierarchical learning model

References

Cassidy, K. W. , & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language, 30,* 348-369.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition, 87*, B23–B34.

Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy, 15,* 125-150.

Landau, B., Smith, L. B., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 5,* 287–312.

Macnamara, J. (1972). The cognitive basis of language learning in infants. *Psychological Review, 79,* 1–13.

Markman, E. M. (1989). *Categorization and naming in children.* Cambridge, MA: MIT Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20,* 121–157.

Smith, L.B., Jones, S.S., Landau, B., Gershkoff-Stowe, L. & Samuelson, L. (2002). Object Name Learning Provides On-the-Job Training for Attention. *Psychological Science, 13,* 13-19.