Simulating Language
3: Frequency learning and regularization

Kenny Smith
kenny.smith@ed.ac.uk

Finishing up on word learning (for now)

Xu, F., & Tenenbaum, J. B. (2007) Word learning as Bayesian Inference. Psychological Review, 114, 245-272

Their task

These are *feps*



Show me all the *feps*











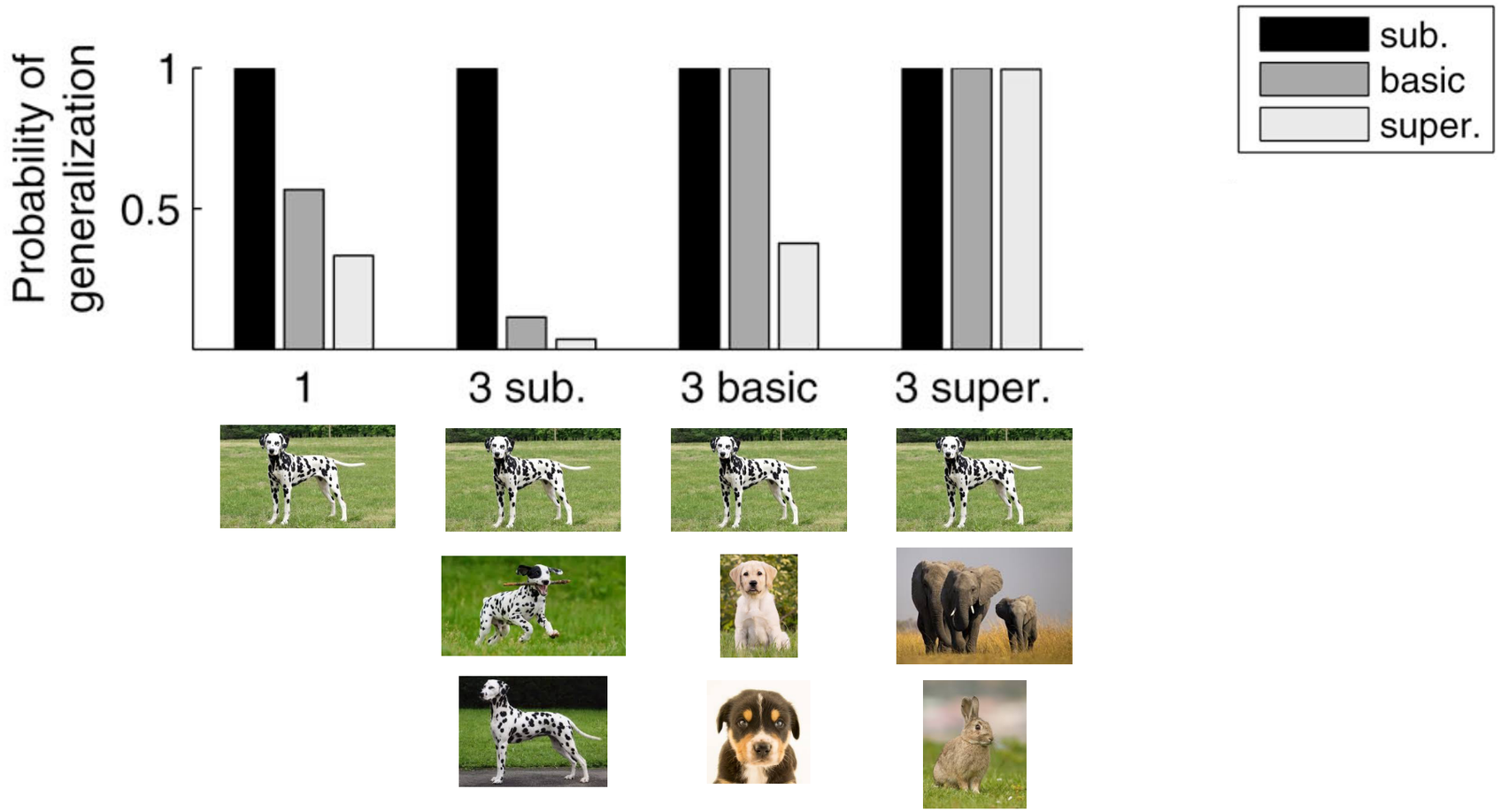







Model predictions

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

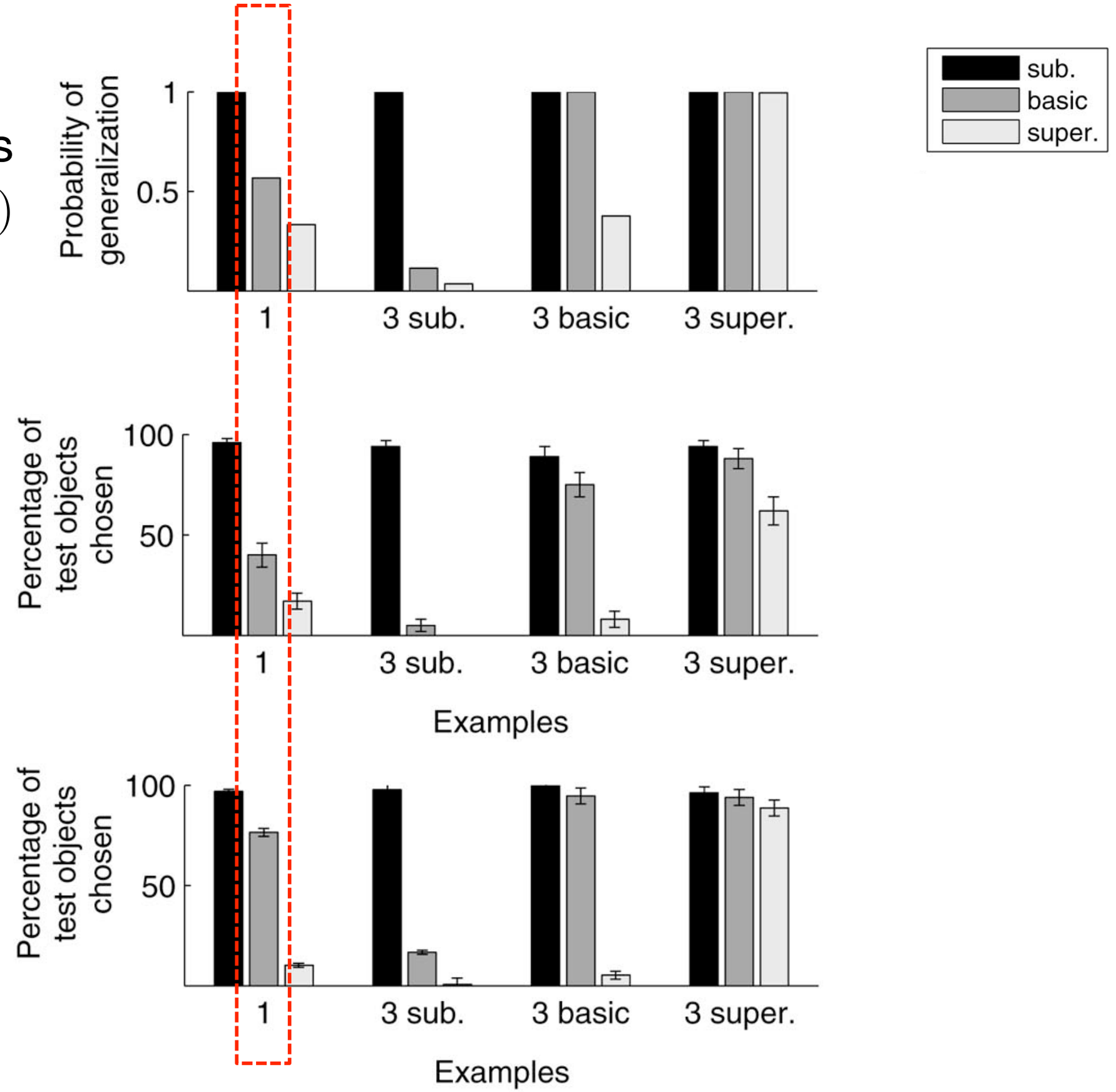


Model predictions

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

Children

Adults



Add a basic-level bias

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

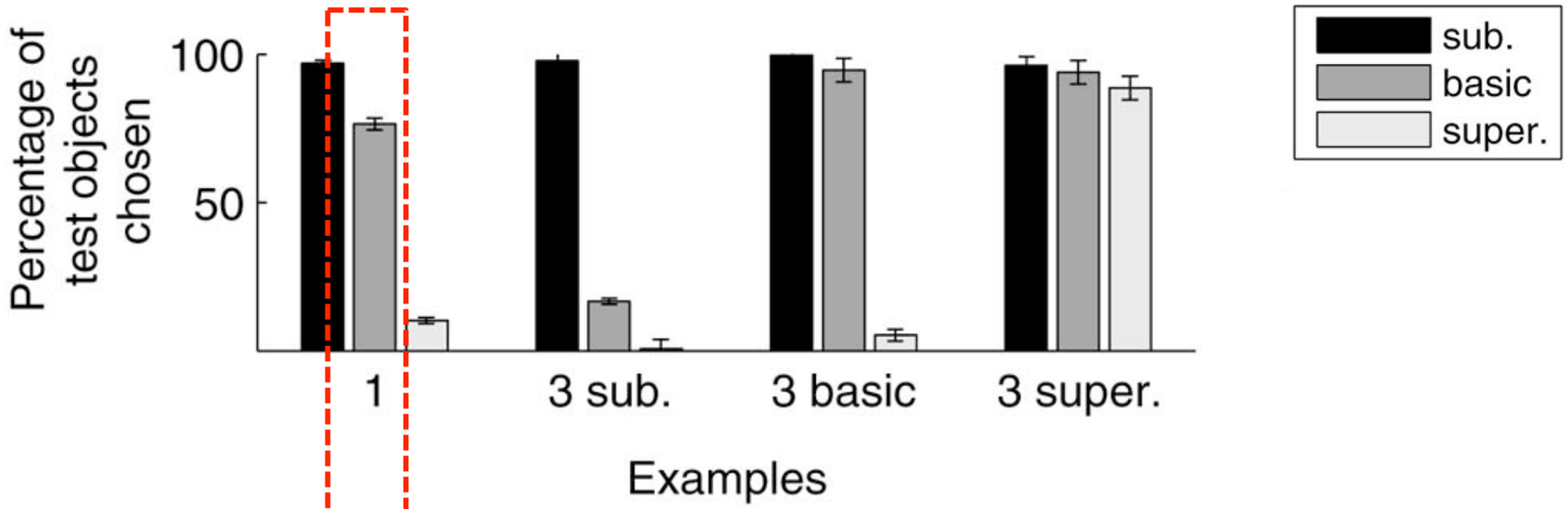
Uniform prior

P(*fep*=dalmatian') = P(*fep*=dog') = P(*fep*=animal')

Prior with a basic-level bias

P(*fep*=dog') > P(*fep*=dalmatian') = P(*fep*=animal')

Adults

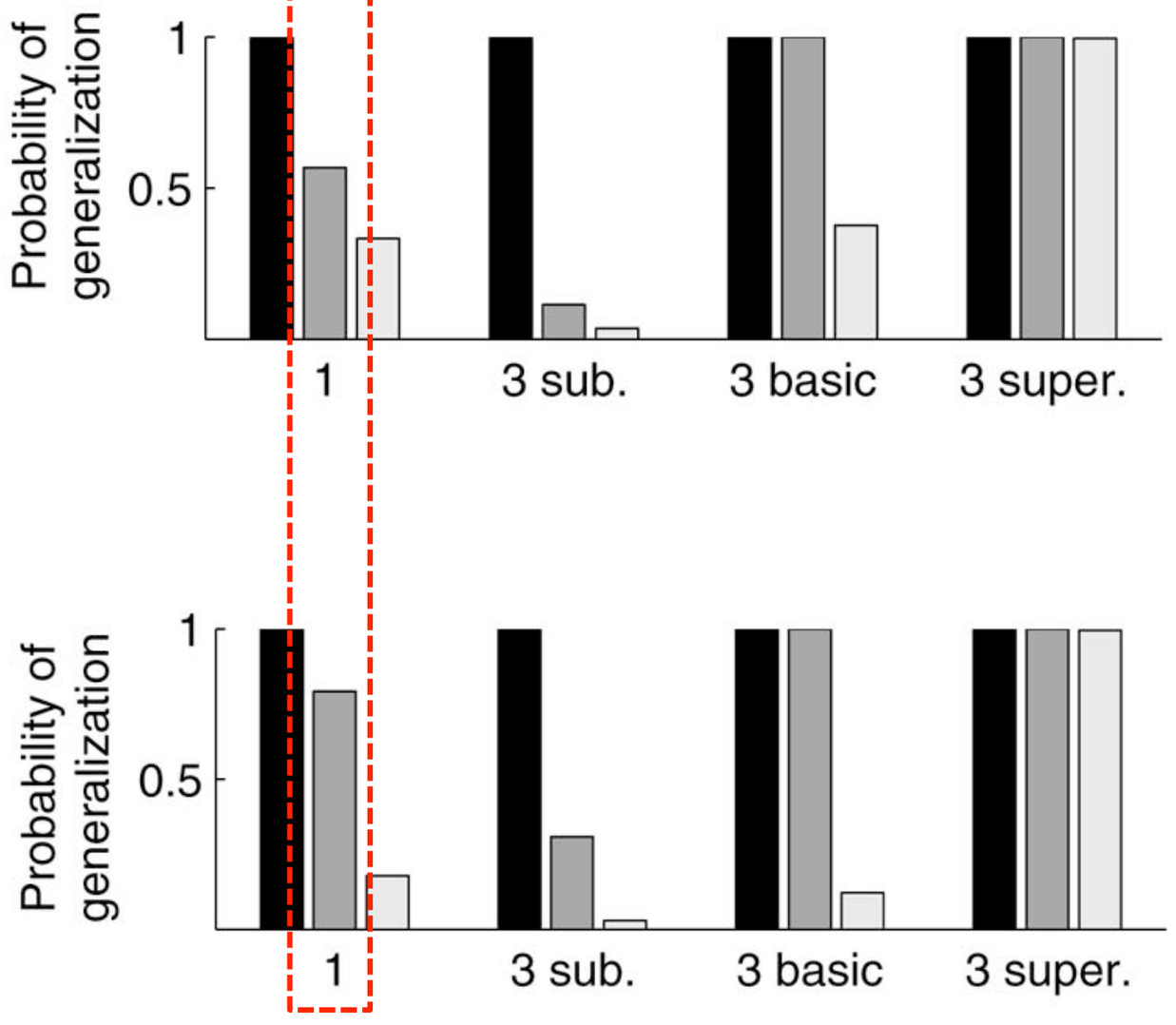


Model without basic-level bias

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

Model **with** basic-level bias

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



Why might adults and children come to this word learning task with different priors?

New topic: Frequency learning and regularization

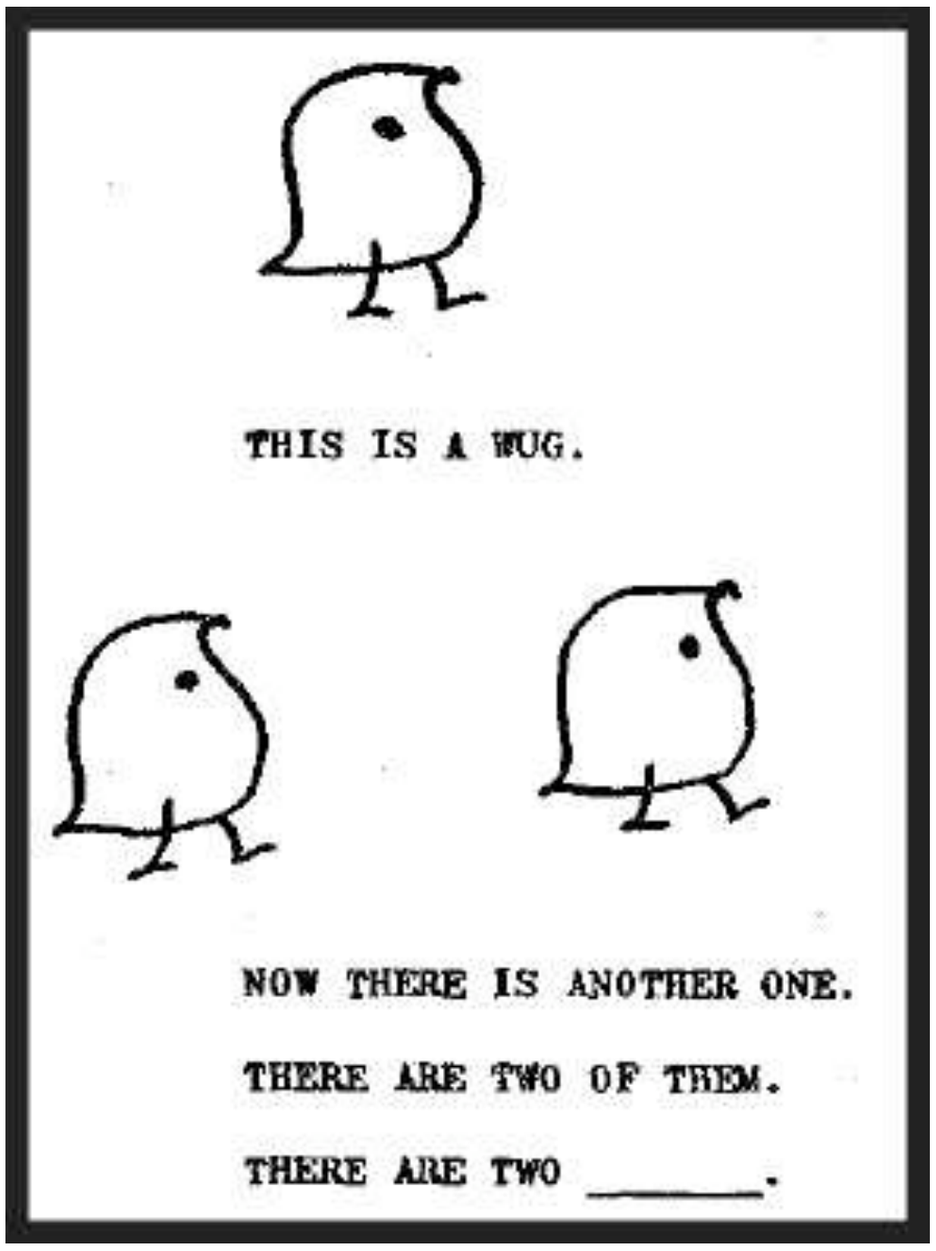
Variation in language

- **An observation:** languages tend to avoid having two or more forms which occur in identical contexts and perform precisely the same functions

- Within individual languages: phonological or sociolinguistic **conditioning** of alternation

- Over time: historical tendency towards analogical levelling

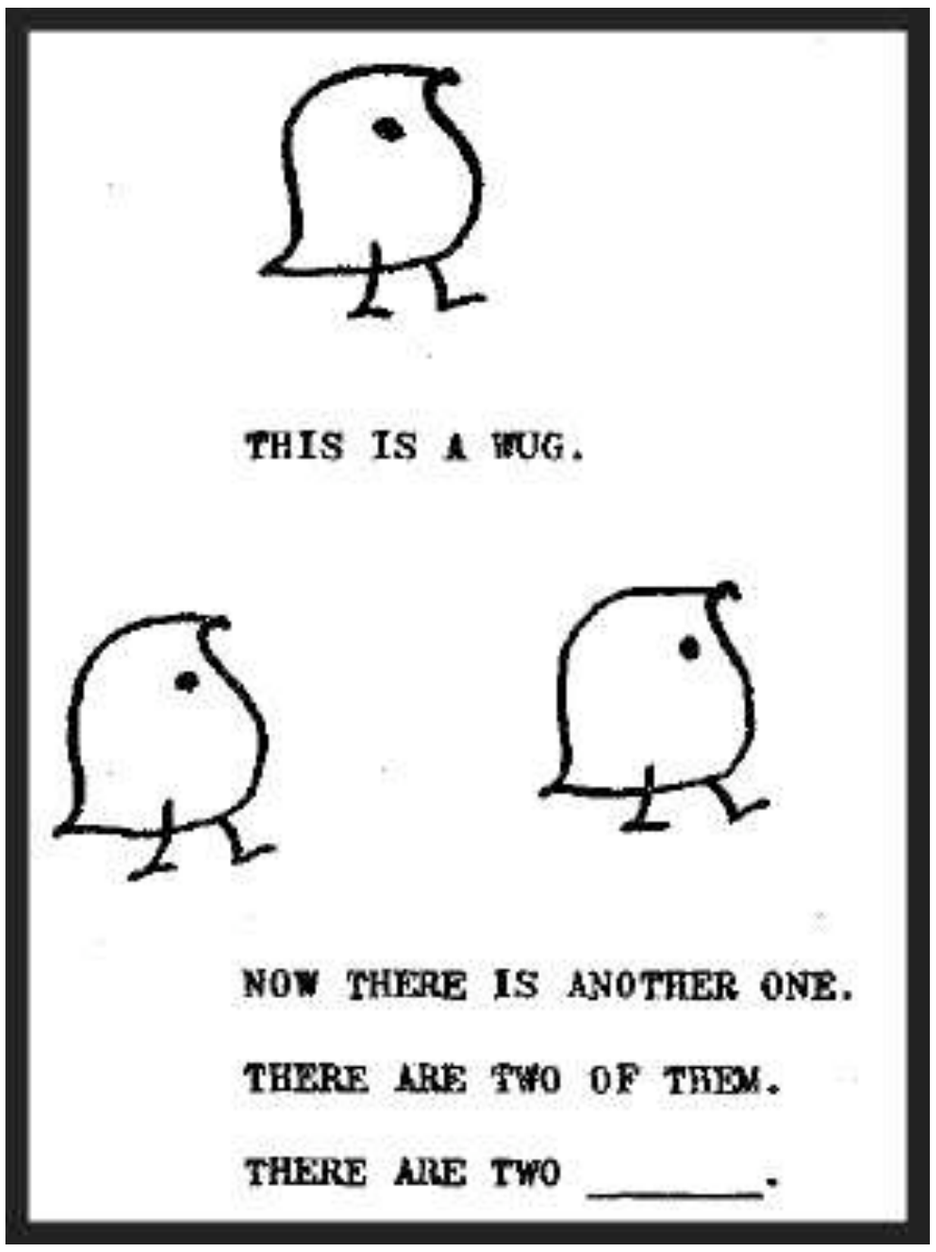
The wug test (Berko, 1958)



- **“wugs”**
- Not “wugen”
  - ox, oxen
- Not “wug”
  - sheep, sheep
- Not “weeg”
  - foot, feet

These ways of marking the plural are relics of older systems which have died out: **loss of variability**

The wug test continued



- Three allomorphs for the regular plural, conditioned on phonology of stem
  - One wug, two /wʌgz/
  - One wup, two /wʌps/
  - One wass, two /wasəz/
- **Conditioning** of variation

Variation in language

- **An observation:** languages tend to avoid having two or more forms which occur in identical contexts and perform precisely the same functions

- Within individual languages: phonological or sociolinguistic conditioning of alternation

- Over time: historical tendency towards analogical levelling

- **During development:** Mutual exclusivity; overregularization of morphological paradigms

Maybe biases in learning, patterns of language change, and the way languages work are all related somehow?

An artificial language learning study

Hudson-Kam & Newport (2005)

- Adults trained and tested on an artificial language
    - 36 nouns, 12 verbs, negation, **2 determiners**
- Multiple training sessions
- Variable (unpredictable) use of ‘determiners’

An artificial language learning study

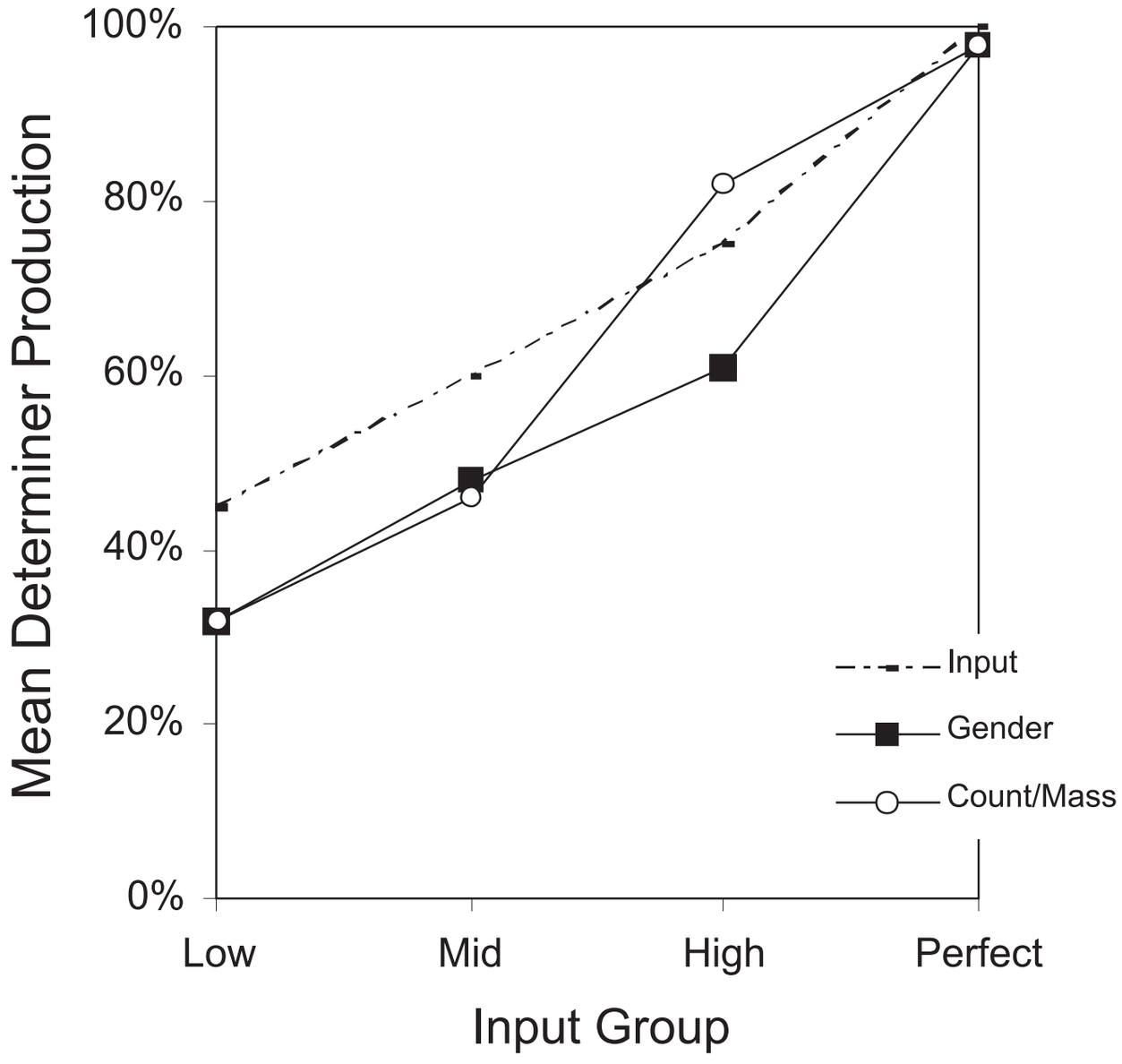
Hudson-Kam & Newport (2005), Experiment 1

- Adults trained and tested on an artificial language
    - 36 nouns, 12 verbs, negation, **2 determiners**
- Multiple training sessions
- Variable (unpredictable) use of ‘determiners’

flern blergen **(ka)** flugat **(ka)**
rams elephant **(Det)** giraffe **(Det)**
“the elephant rams the giraffe”

Adults **probability match**

Probability matching: if trained on variable input, produce variable output, matching the input frequencies.



Adults **probability match**

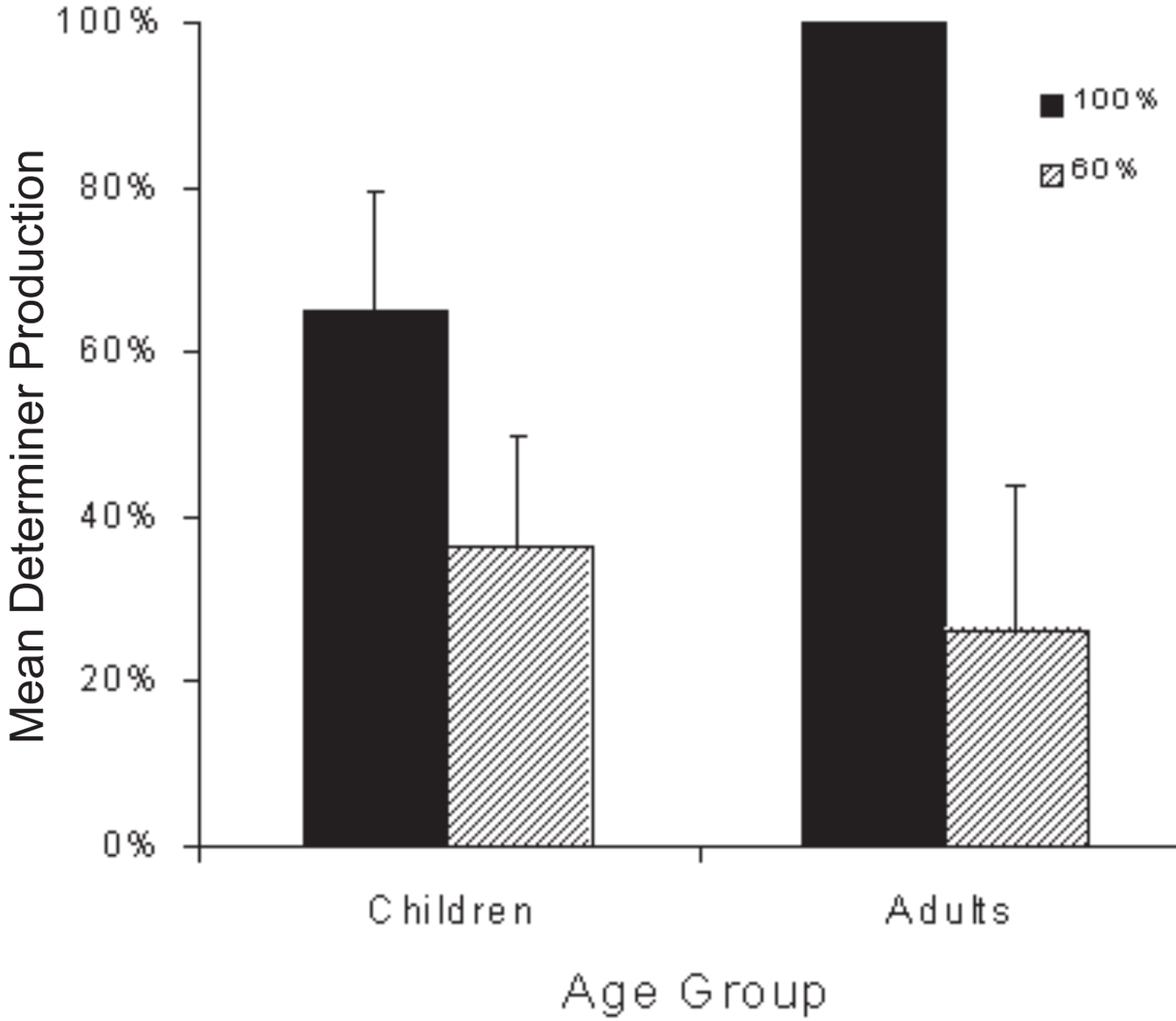
Probability matching: if trained on variable input, produce variable output, matching the input frequencies.

	Production Type		
Input Group	*Systematic User*	*Systematic Non-User*	*Variable User*
100	100.0	0.0	0.0
75	11.1	11.1	77.8
60	0.0	25.0	75.0
45	0.0	0.0	100.0

Hudson-Kam & Newport (2005), Experiment 2

- Adults **and children (age 6;4)** trained and tested on an artificial language
    - 12 nouns, 4 verbs, **1 determiner**
- Multiple training sessions
- Variable (unpredictable) use of the determiner

Kids are more variable?



Kids (somewhat) more likely to **regularize**

Regularization: if trained on variable input, produce non-variable output.

	Production Type				
Input Group	*Systematic User*	*Systematic Nonuser*	*Systematic Other*	*Systematic Total*	*Variable User*
Children					
100%	50.0	25.0	12.5	87.5	12.5
60%	14.3	57.0	0.0	71.3	28.6
Adults					
100%	100.0	0.0	0.0	100.0	0.0
60%	0.0	50.0	0.0	50.0	50.0

What's going on here?

- Do adults have the 'wrong' bias to explain how language is, how language changes?

- Do kids have different biases (i.e. a different prior)?

- Or do we just have bad intuitions about how a biased learner should behave?

- We need a model

  - Beta-binomial model from Reali & Griffiths (2009) - we'll get to their paper next week

The model in a nutshell

- Let’s simplify: one grammatical function, two words which could mark it
  - word 0, word 1
- The learner gets some data
  - word 0, word 0, word 1, word 1, word 0, ...
  - ∅, ∅, ka, ka, ∅, ...
- And has to infer how often it should use each word
  - “I will use word 0 60% of the time, and word 1 40% of the time”
  - “I will use word 1 40% of the time”
  - $\theta = 0.4$

A little more detail

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

- The learner gets some data, d
  - word 0, word 0, word 1, word 1, word 0, ...
- And has to infer how often it should use each word, based on that data
  - θ
- The learner will consider several possible hypotheses about θ
  - Is word 1 being used 5% of the time? 15%? 25%? ...
  - θ = 0.05? θ = 0.15? θ = 0.25? ...
- The learner will use Bayesian inference to decide what θ is

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

The likelihood

- Let’s say that the probability of using word 1 is 0.5 - both words are equally likely to be used
    - $\theta = 0.5$
- Let’s say your data consists of a single item: a single occurrence of word 1
    - $d = [1]$
- What is the likelihood of this data, given that $\theta = 0.5$? i.e. what is $p(d = [1] \mid \theta = 0.5)$?
- What is $p(d = [1,1,1] \mid \theta = 0.5)$?
- What is $p(d = [1,1,1] \mid \theta = 0.1)$?

The likelihood: summary

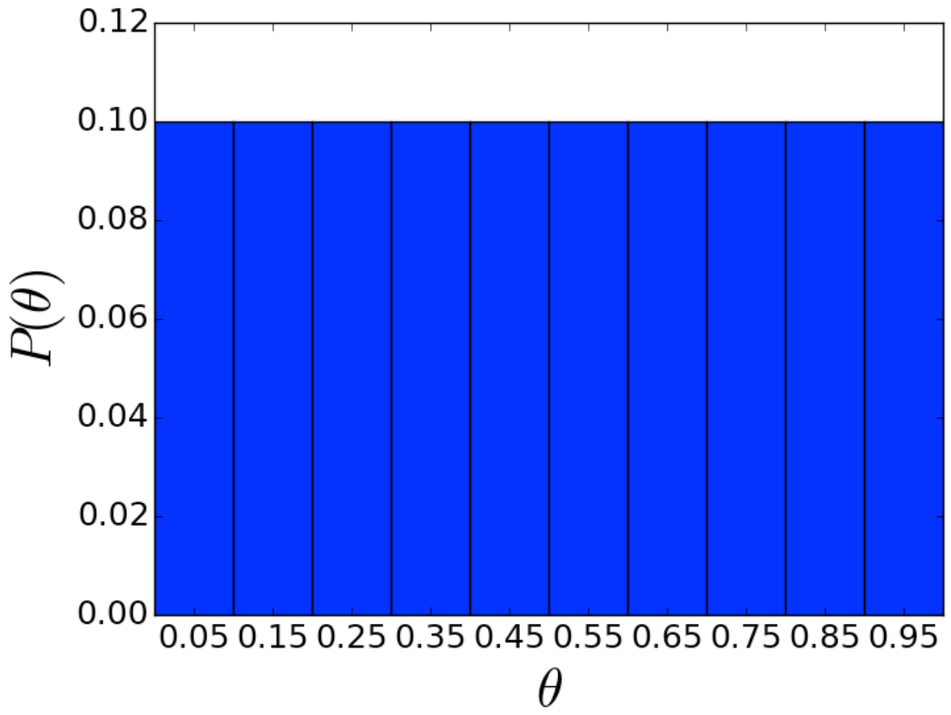
- When θ is high, data containing lots of word 1 is very likely
- When θ is around 0.5, data containing lots of word 1 is less likely
    - A mix of 1s and 0s is more likely
- When θ is low, data containing lots of word 1 is very unlikely
    - Lots of word 0 is more likely

The prior

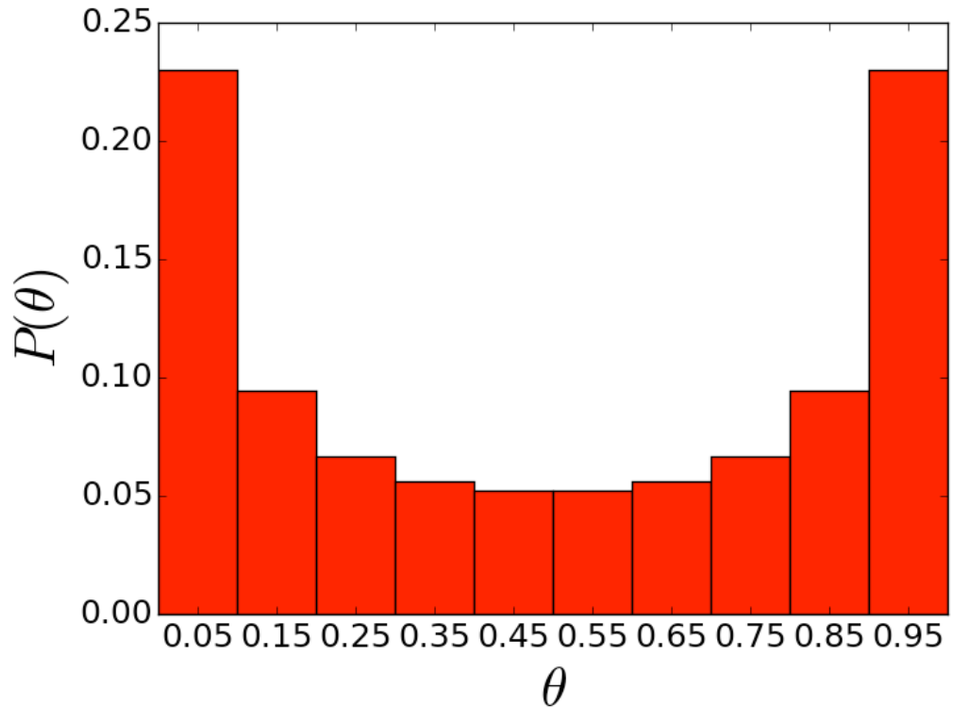
- Let’s say our learner considers 10 possible values of θ
    - 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95
- Our prior says, for each possible value of θ, how likely our learner thinks it is, before they have seen any data
    - High prior probability for a given value of θ means, before seeing any data, the learner thinks that value is likely
    - Low prior probability for a given value of θ means, a priori, the learner thinks that value is unlikely

Which of these possible priors would be a good model for an **unbiased learner**, who thinks each possible value of θ is equally probable a priori?

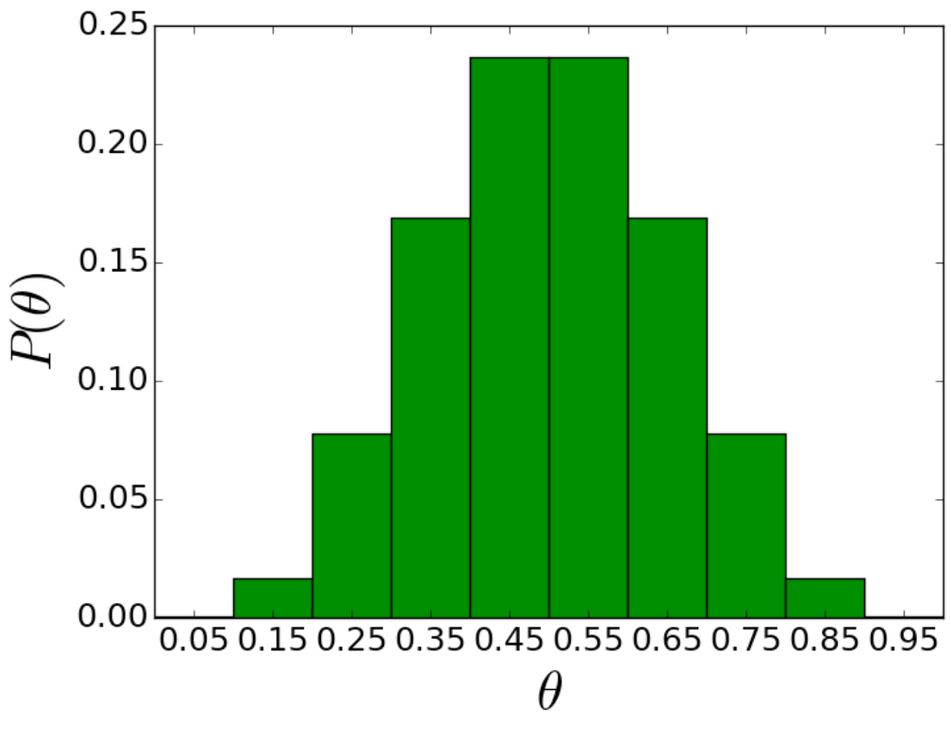
A



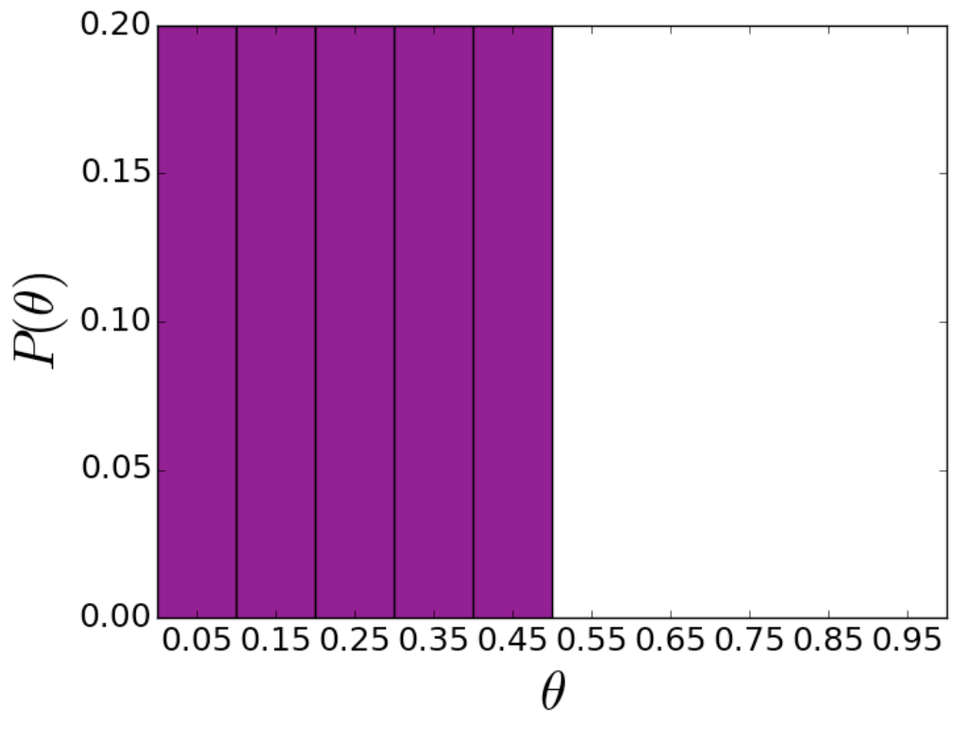
B



C

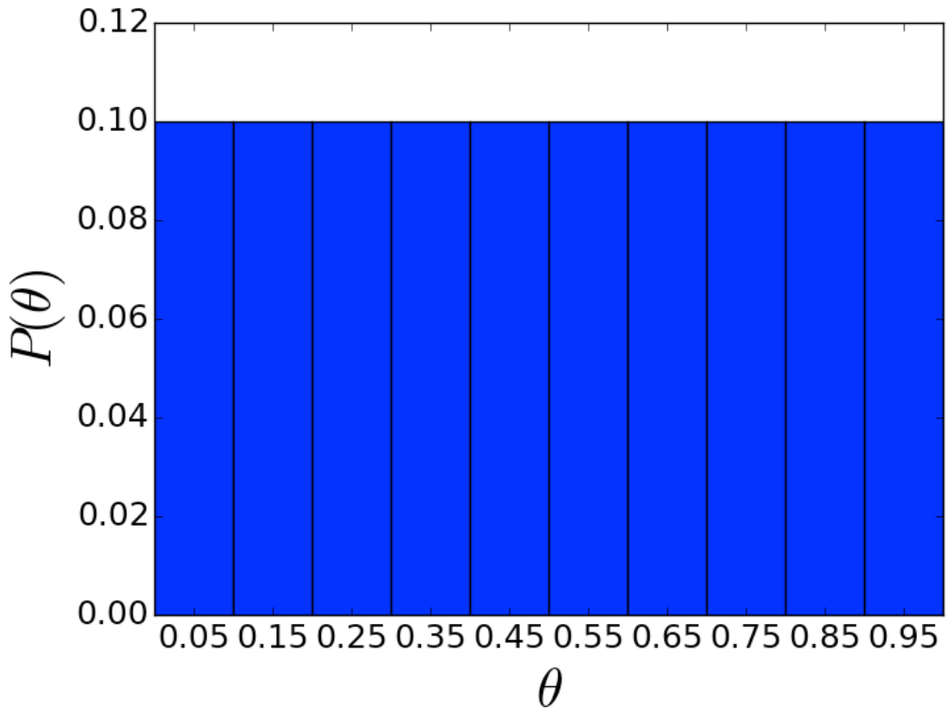


D

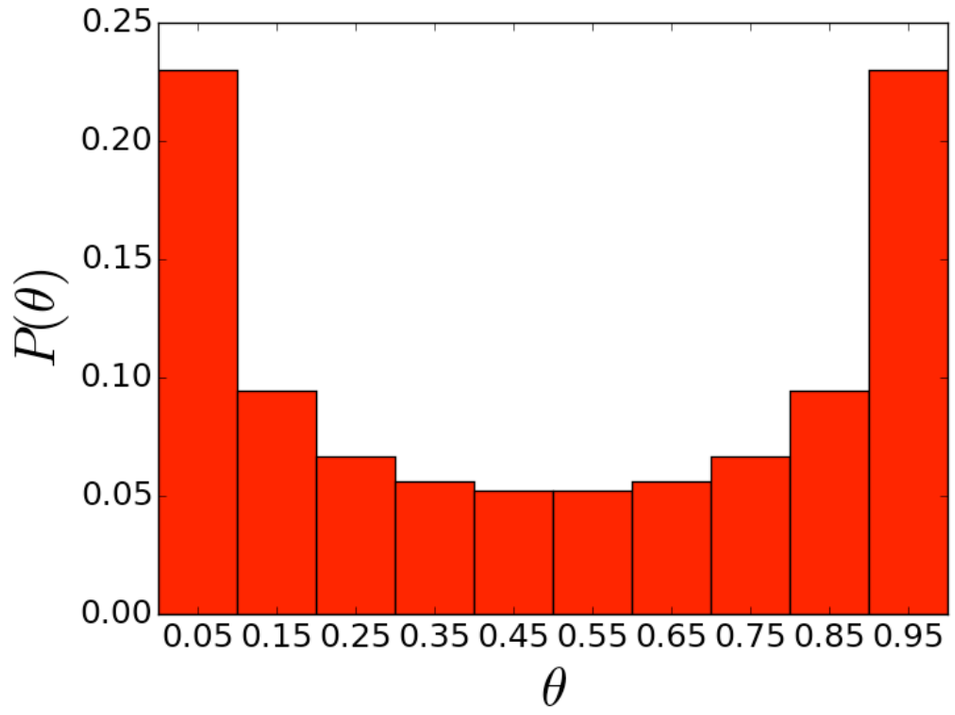


Which of these possible priors would be a good model for a **biased** learner, who thinks **each word should be used roughly equally often**?

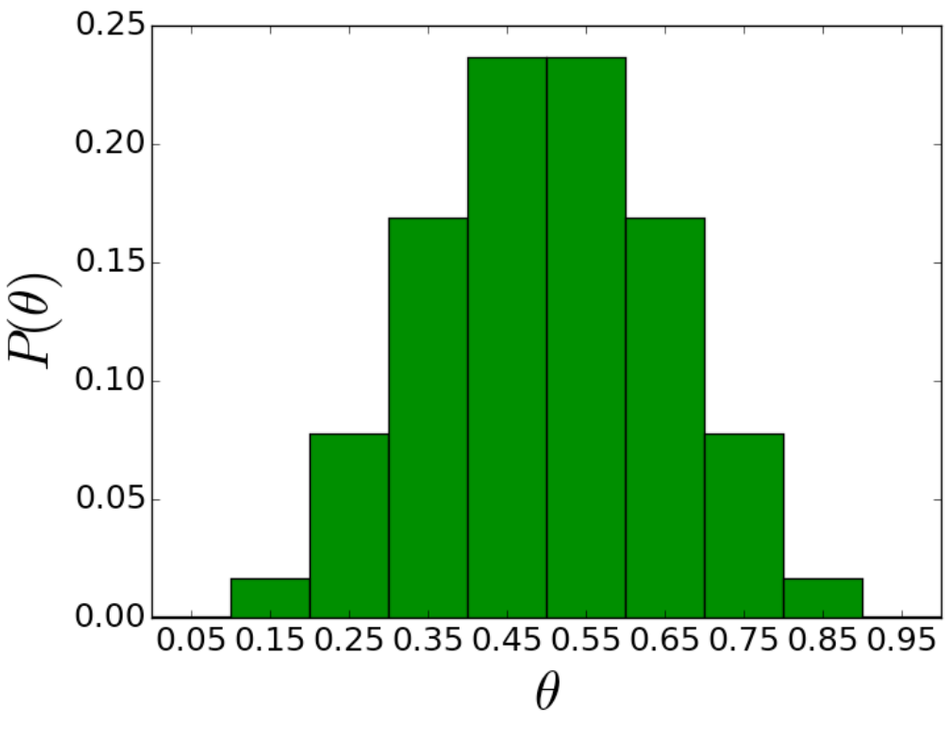
A



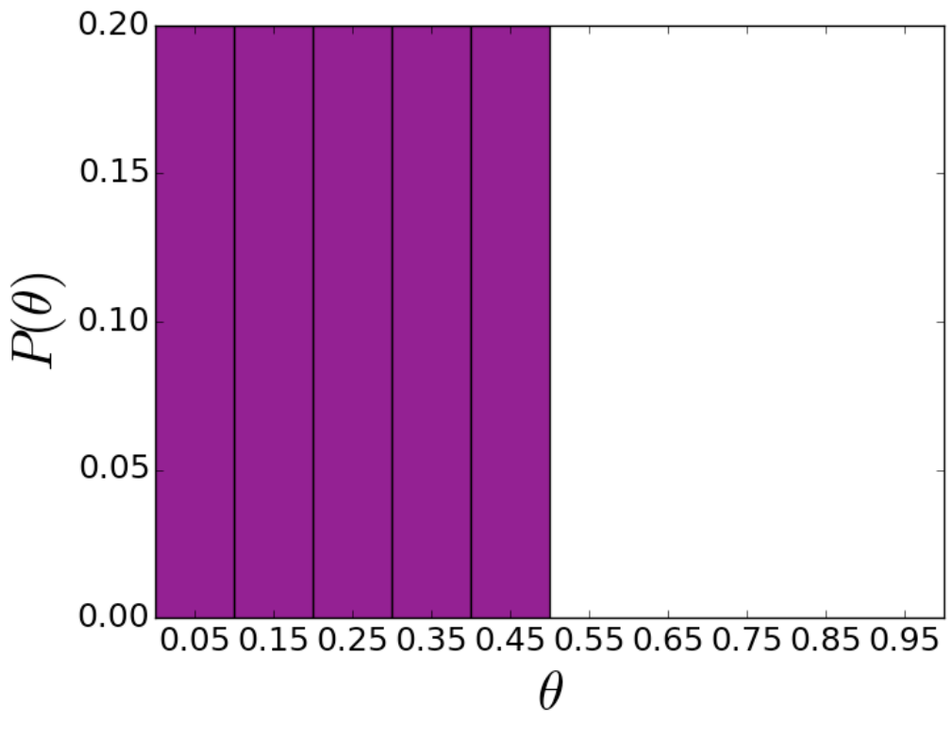
B



C

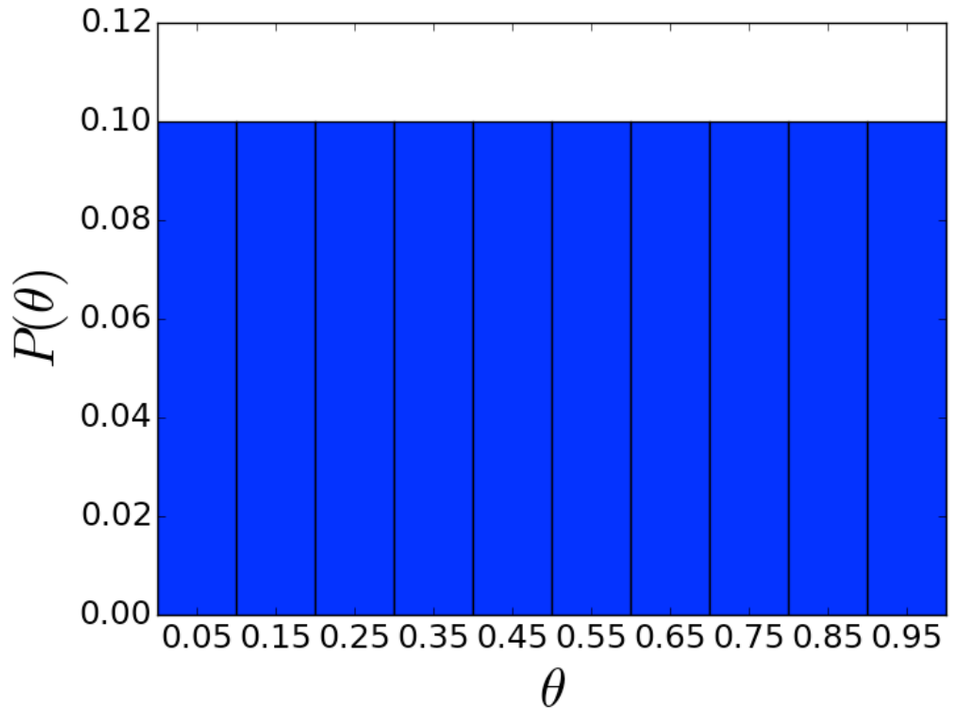


D

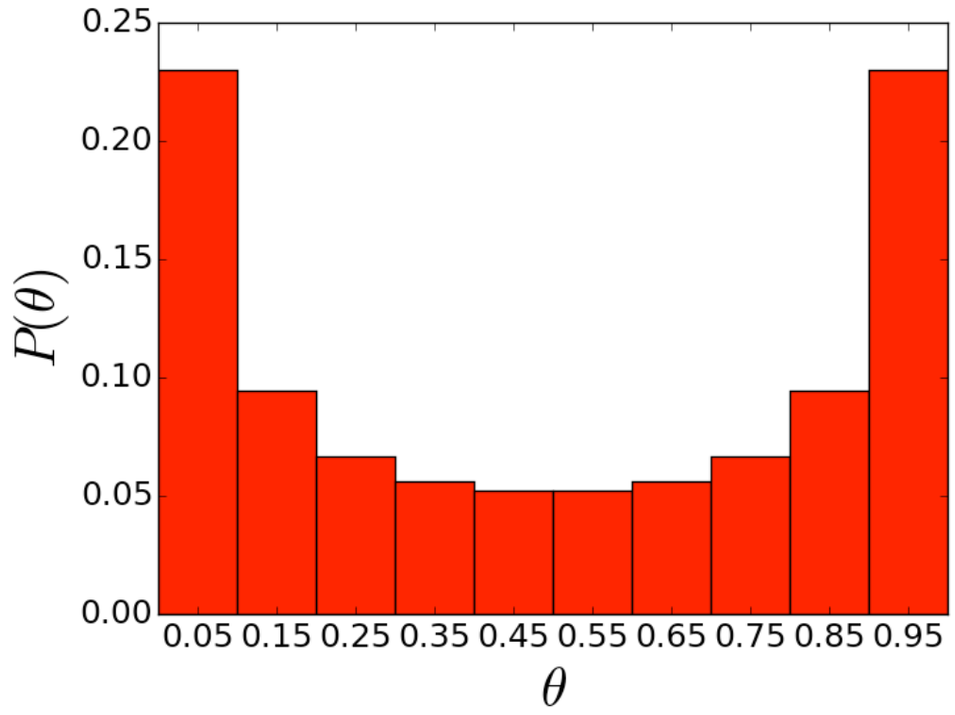


Which of these possible priors would be a good model for a **biased** learner, who thinks **only one word should be used** (but isn't sure if it should be word 0 or word 1)?

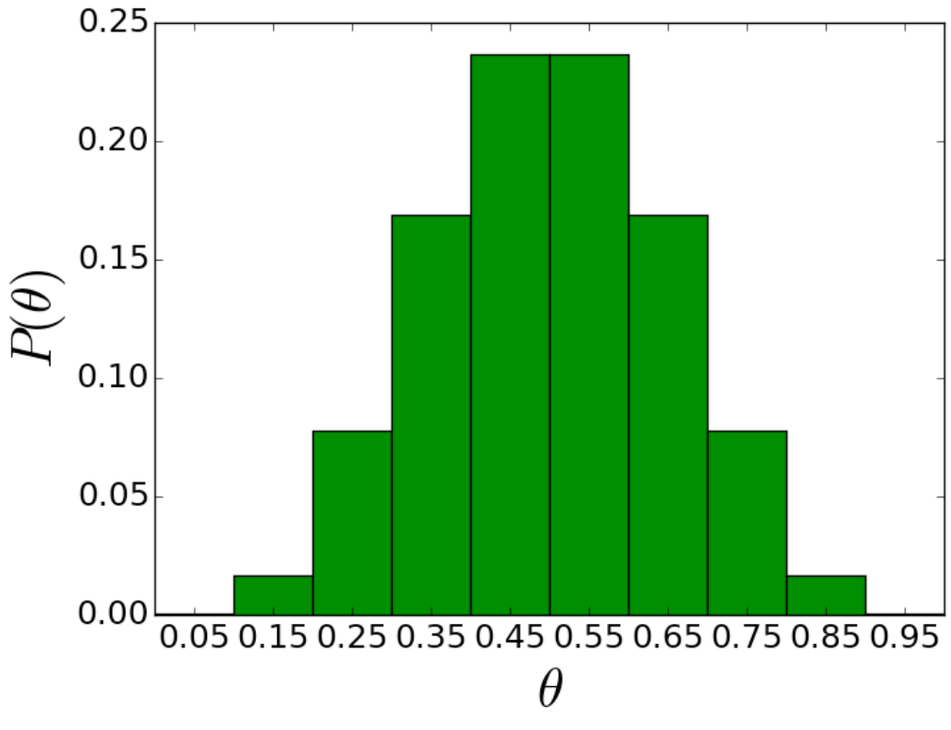
A



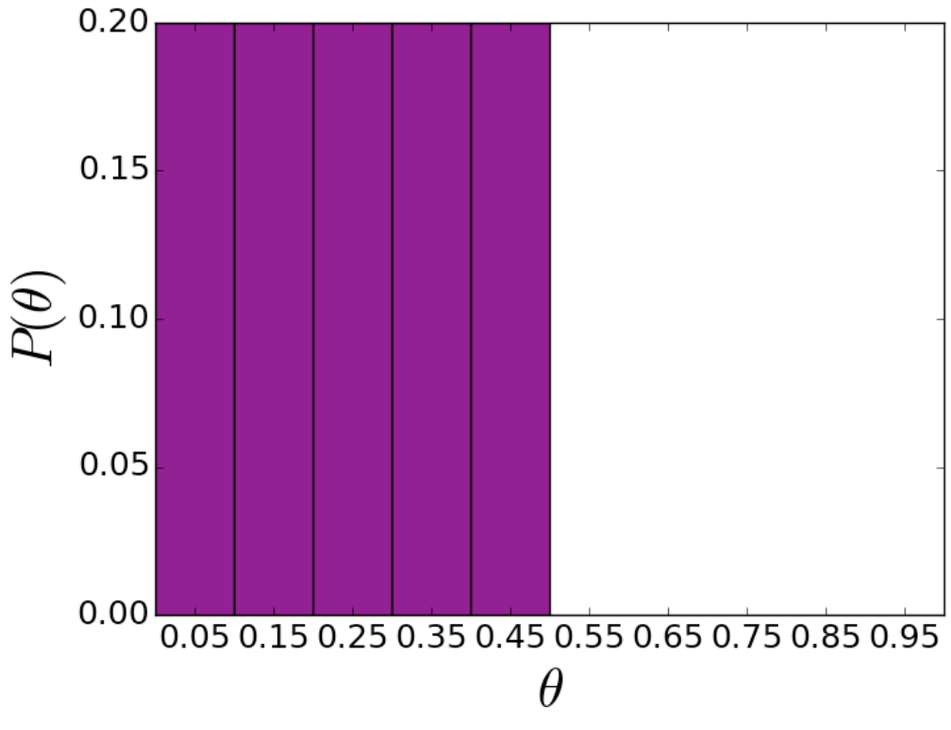
B



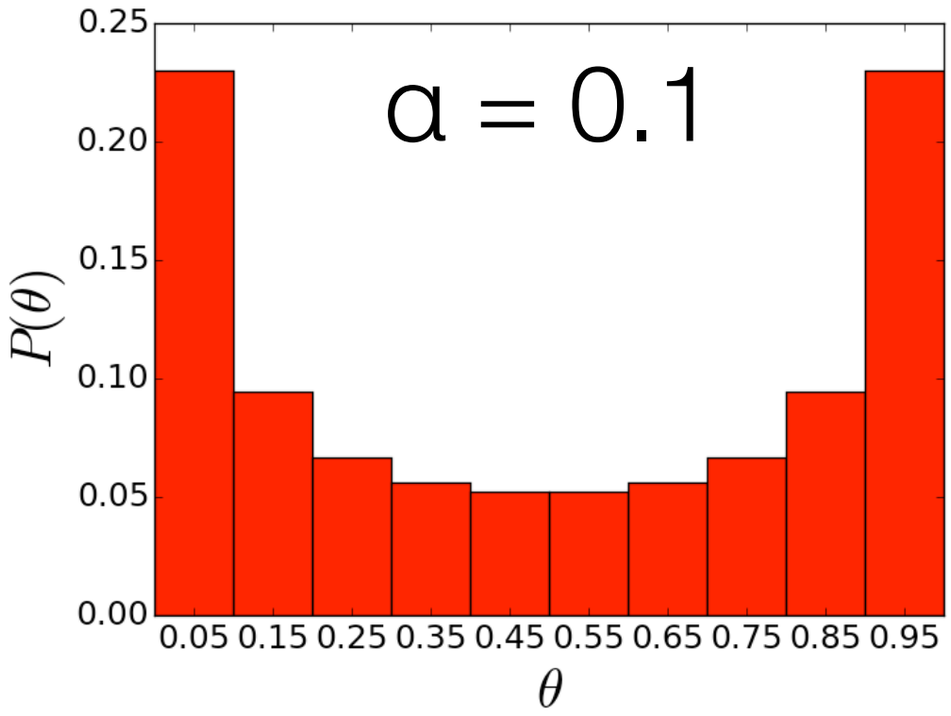
C

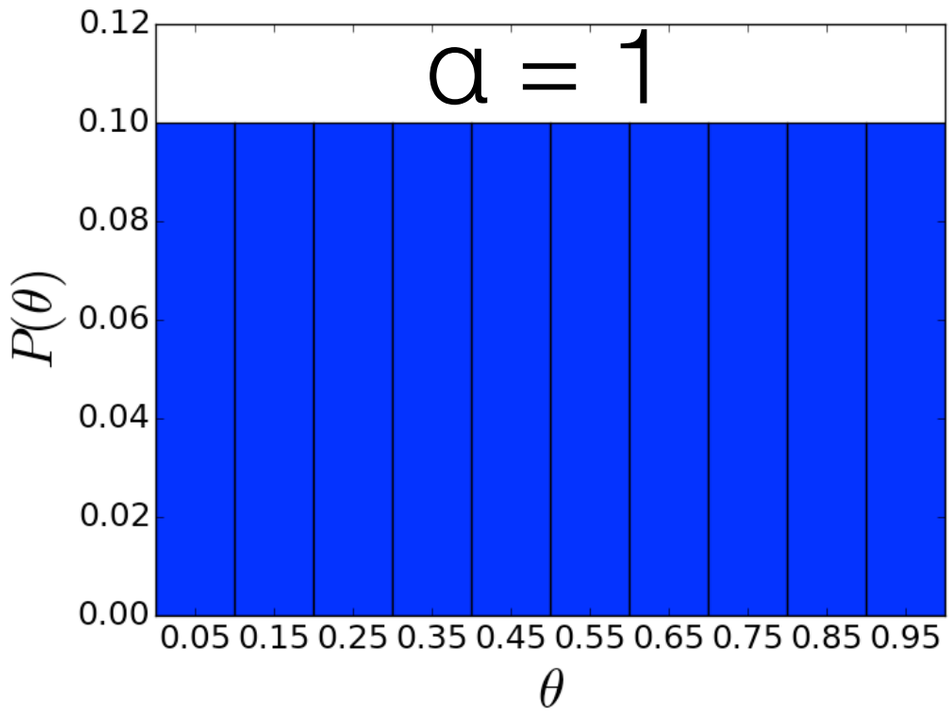


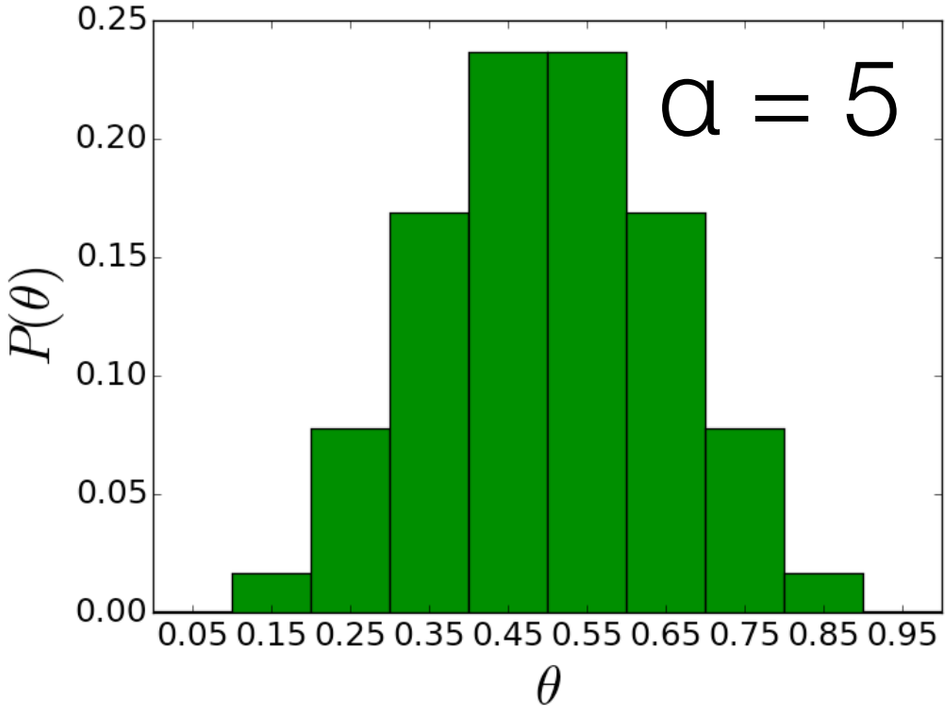
D



Our prior: the (symmetrical) beta distribution

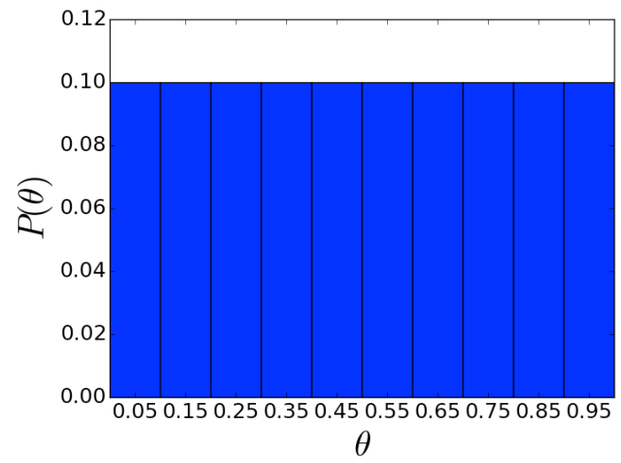






Putting it together

- Let's say our learner considers 10 possible values of θ, i.e. our hypothesis space looks like this: 0.05, 0.15, 0.25, … 0.75, 0.85, 0.95



- They have a **uniform prior**
- And they have some data: d = [1,1]
- We can calculate the posterior probability for each possible value of θ
- This gives us a **posterior probability distribution**

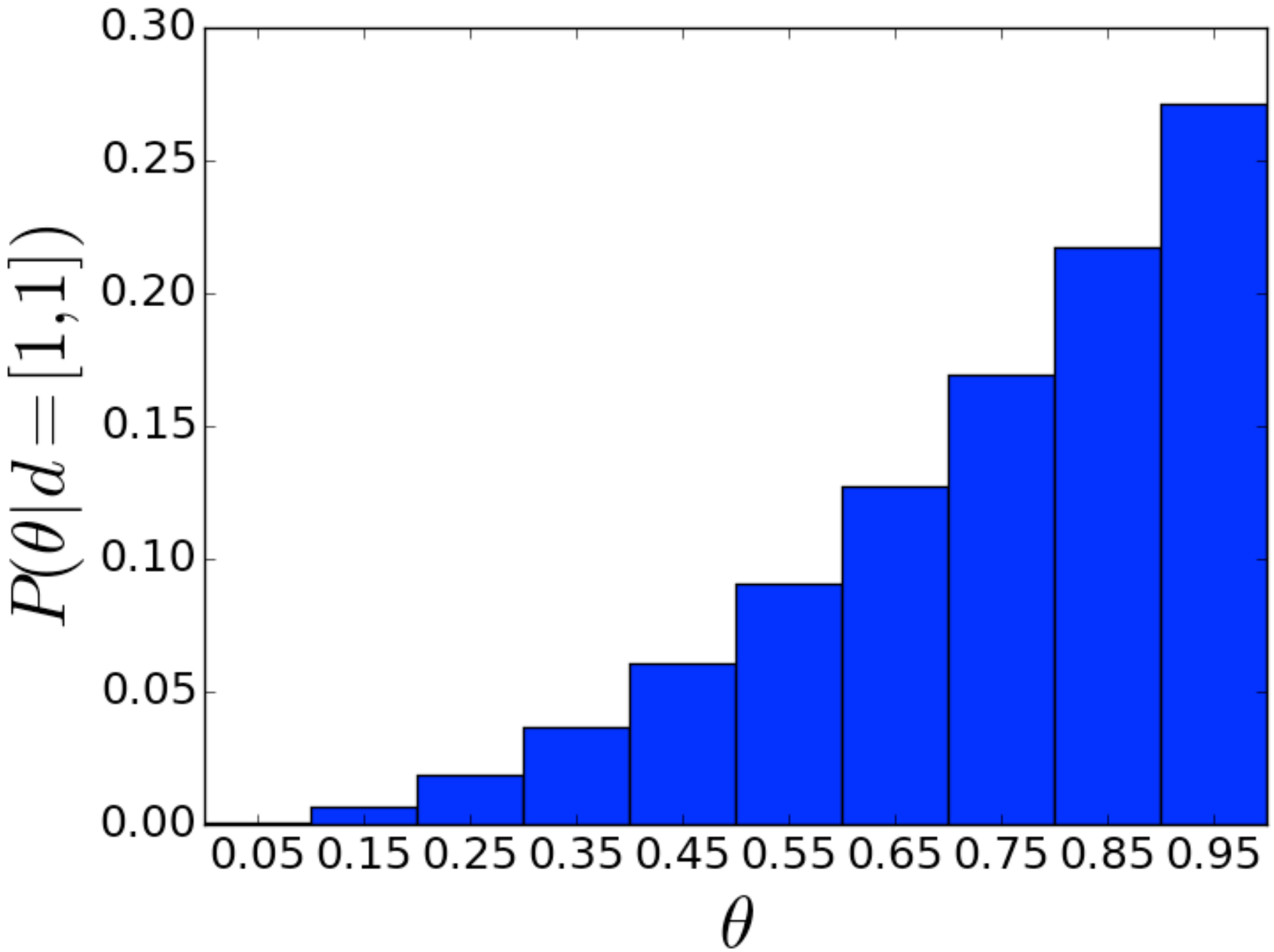
$$P(\theta|d) \propto P(d|\theta)P(\theta)$$

Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

- Uniform prior, d=[1,1]
- Consider just θ=0.25 and θ=0.75.
  - Which has higher posterior probability?
  - How much higher?

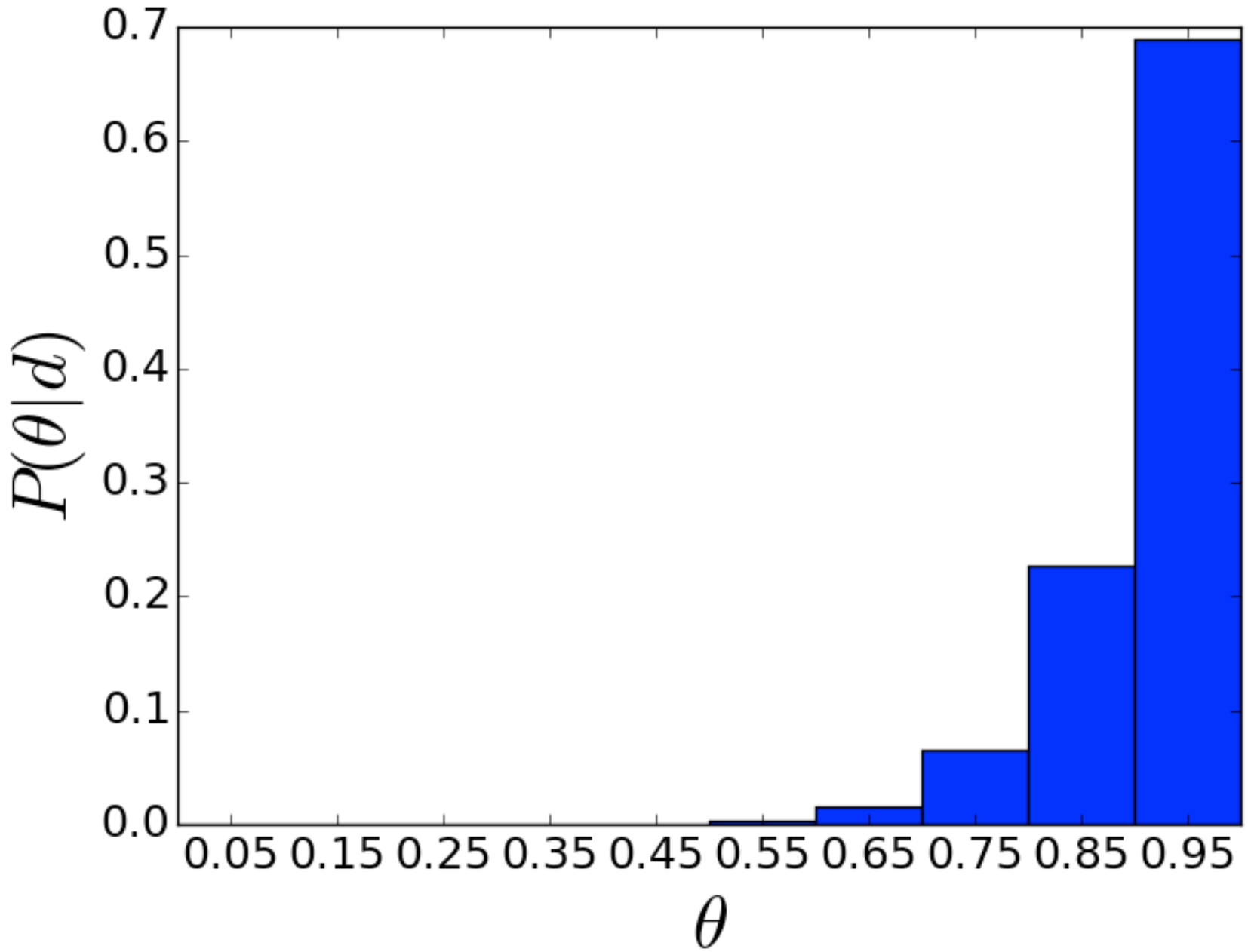
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

- Uniform prior, d=[1,1]



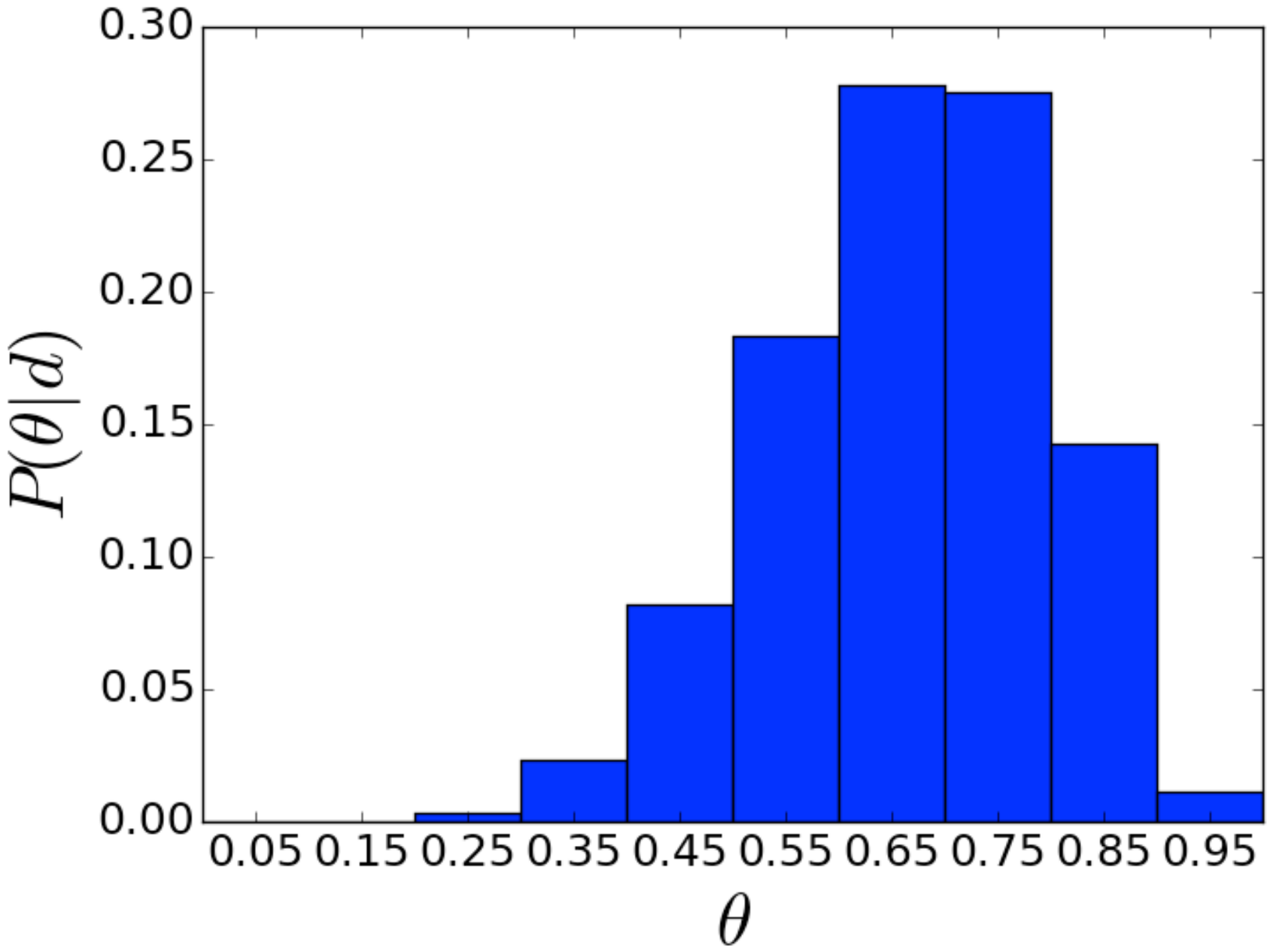
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

- Uniform prior, d=[1,1,1,1,1,1,1,1,1,1]



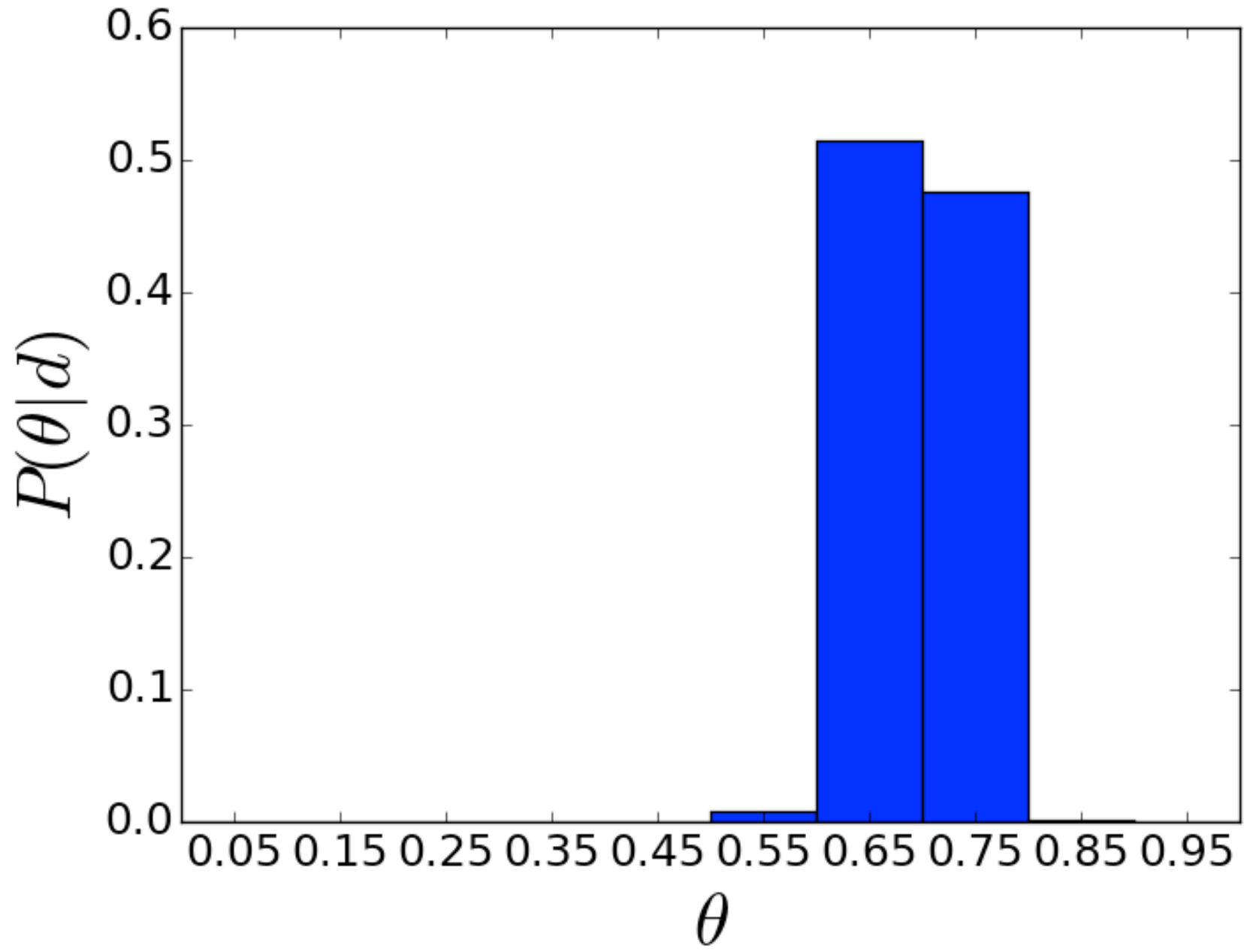
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

- Uniform prior, d=[1,1,1,1,1,1,1,0,0,0]



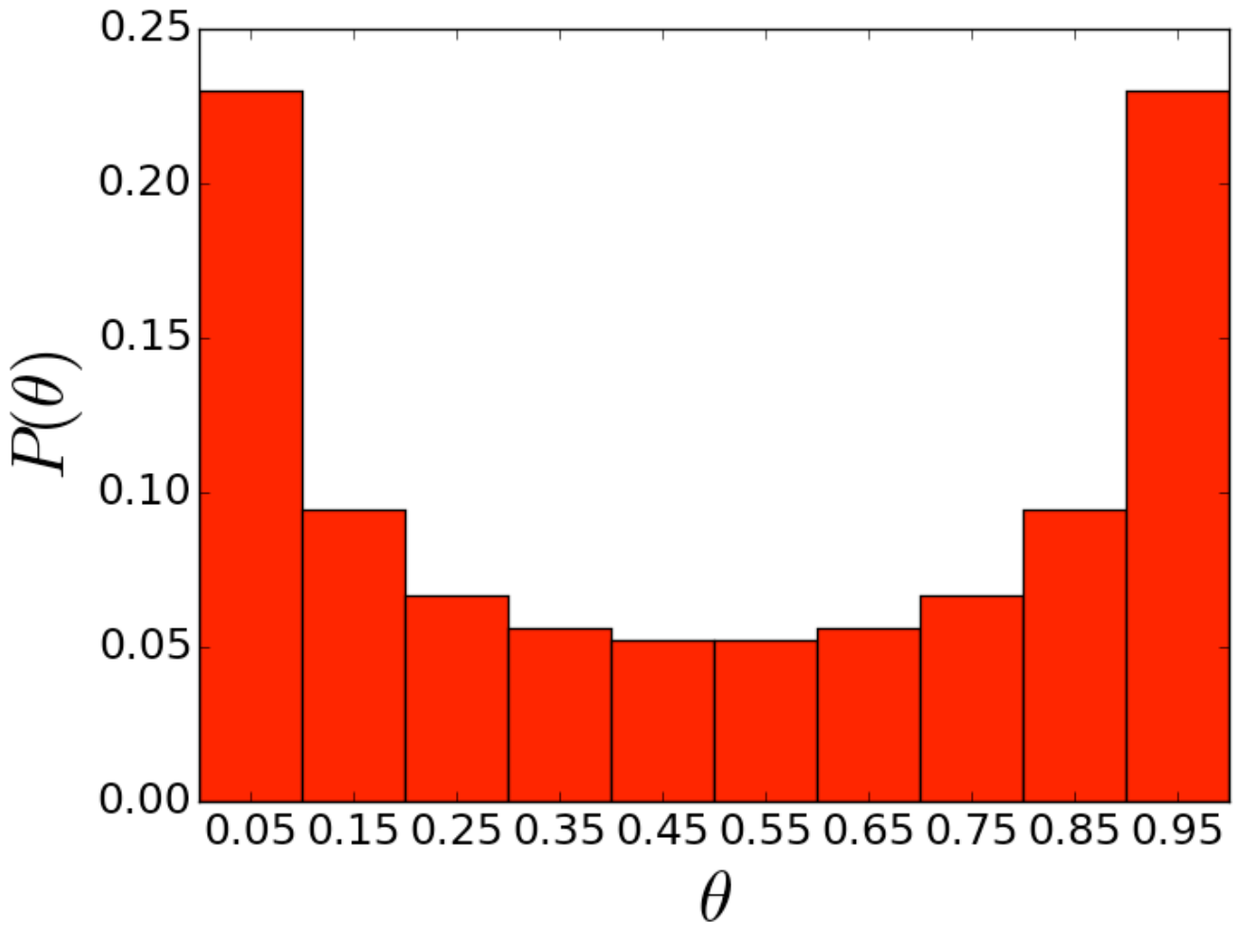
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

- Uniform prior, d=[70 occurrences of word 1, 30 of word 0]



Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

- What happens if we plug in a prior favouring regularity?
- Becomes quite hard to guess: let's run the model!



Coming up next!

- This week's lab: a simple Bayesian model of frequency learning
    - Play around with amount of data and the prior
    - See if you can get probability matching and/or regularization behaviours
- Next week: extending this model to **iterated learning**
    - What happens when learners learn from other learners?

References

Berko, J. (1958). The Child's Learning of English Morphology. *Word, 14,* 150–177.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1,* 151–195.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition, 111,* 317–328.