Simulating Language
16: Iterated Bayesian Learning

Simon Kirby

simon@ling.ed.ac.uk

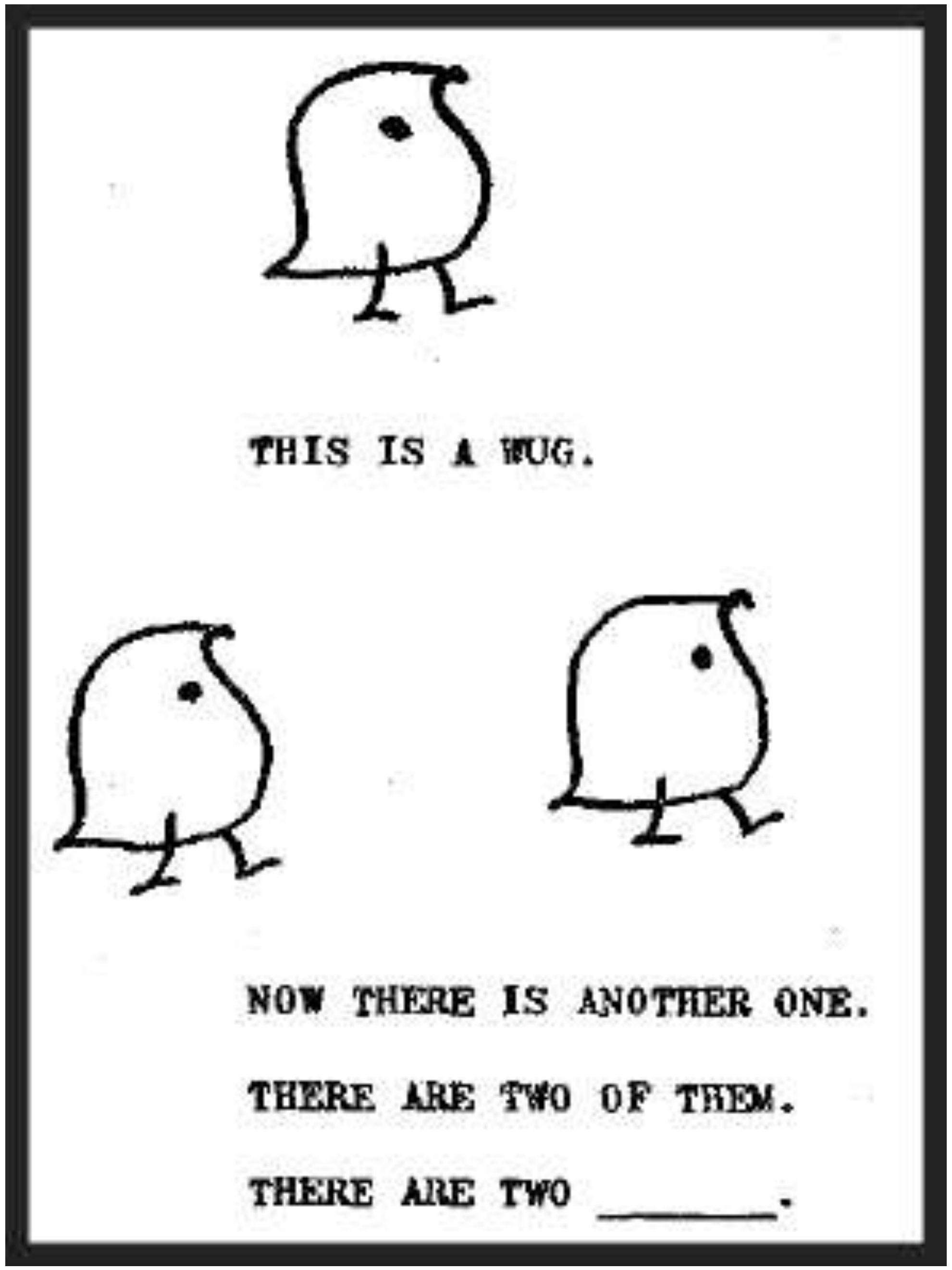
Variation in language

- **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

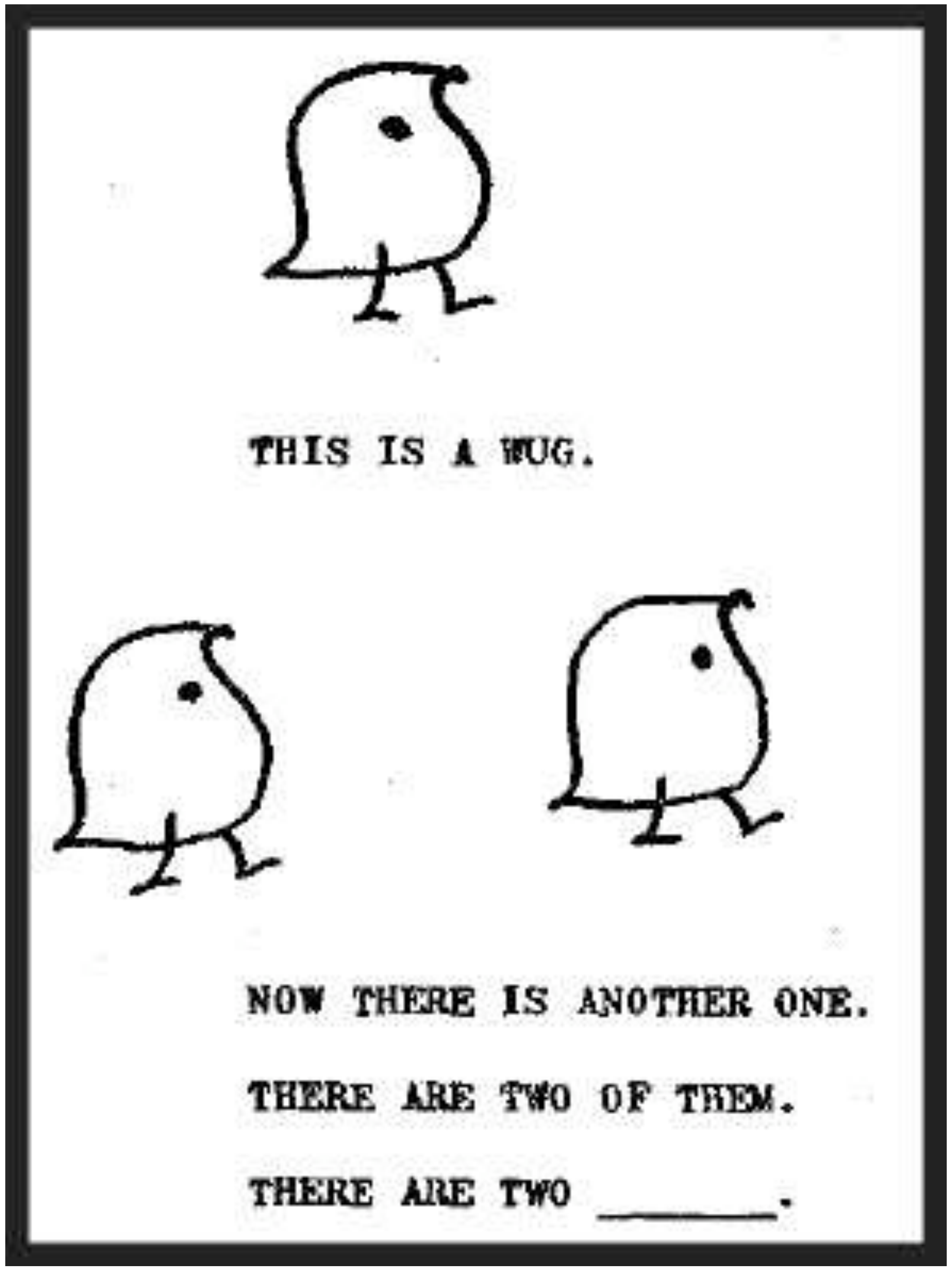
The wug test



- **“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

A prediction about the bias of learners

- Languages tend not to exhibit free (unpredictable, unconditioned) variation
- Languages are transmitted via iterated learning, and should reflect the biases of learners
- We already know that child learners are biased against ‘variation’ in the lexicon (synonymy, Mutual Exclusivity)
- This kind of learning bias is probably pretty widespread, right?

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)

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

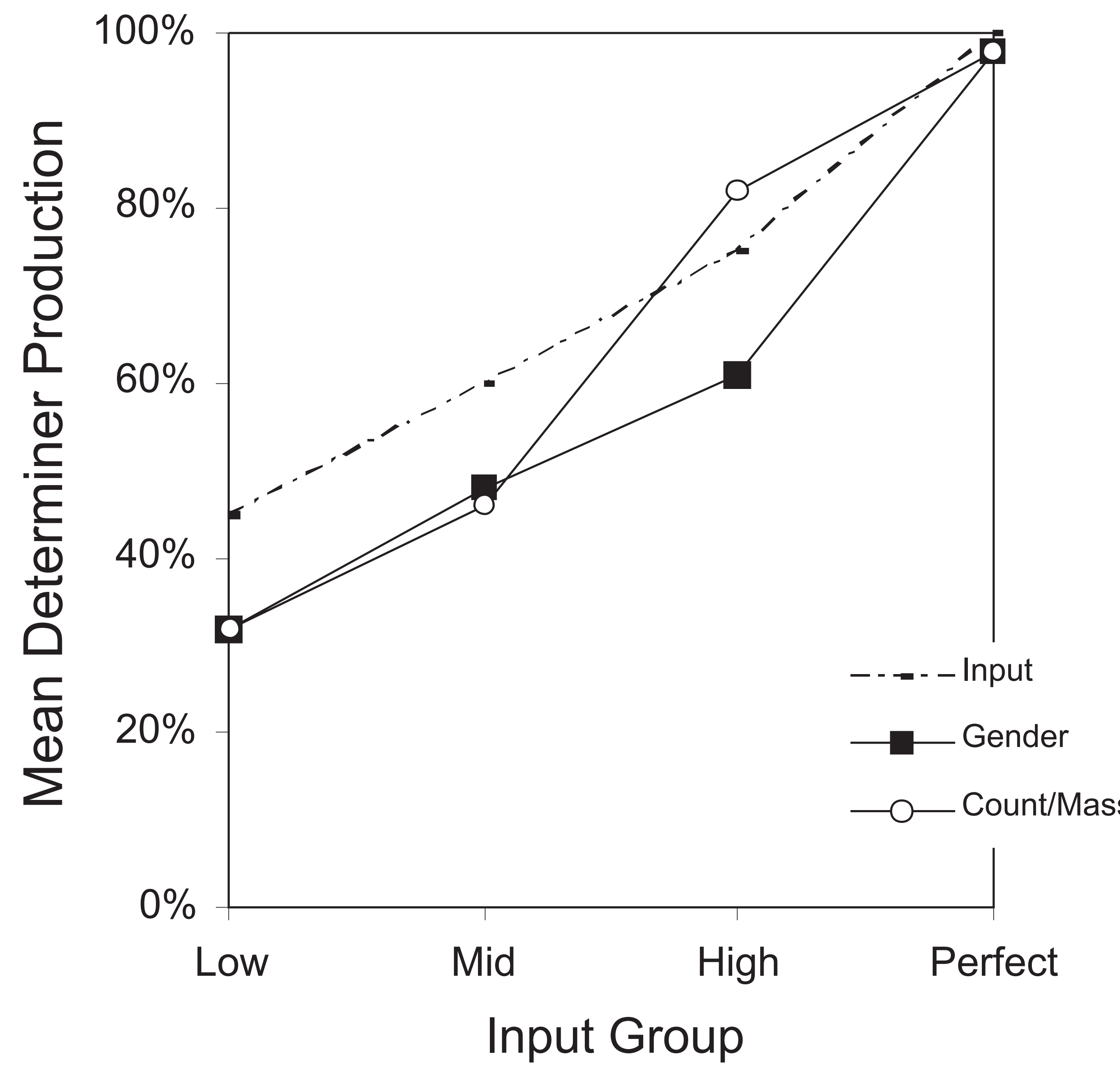
flern	blergen	**(ka)**	flugat	**(ka)**
rams	elephant	**(Det)**	giraffe	**(Det)**

"the elephant rams the giraffe"

Adults **probability match**

- If trained on variable input, produce variable output

- Does this mean they have the 'wrong' bias to explain how language is?
- Or do we just have bad intuitions about how a biased learner should behave?
- We need a model
  - Reali & Griffiths (2009)



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 $\theta = 0.5$?
  - 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

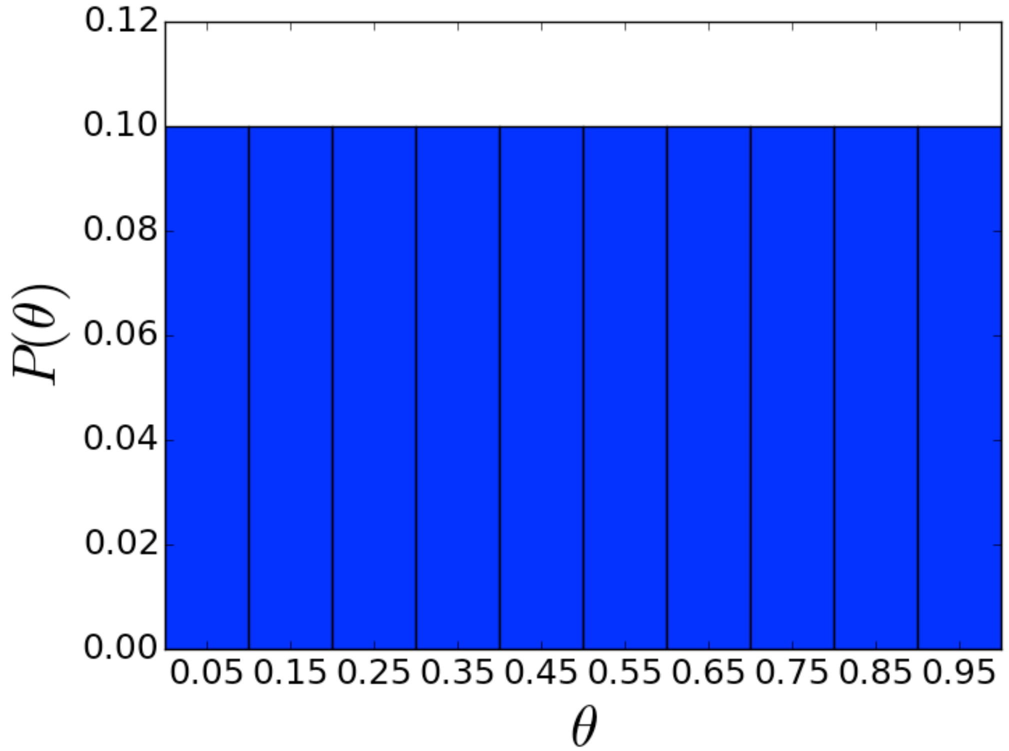
- When θ is high, data containing lots of word 1 is very likely
- When θ is around 0.5, data containing lots of word 1 is not that likely
  - A mix of 1s and 0s is more likely
- When θ is low, data containing lots of word 1 is very unlikely
  - Lots of word 0 is more likely

The prior

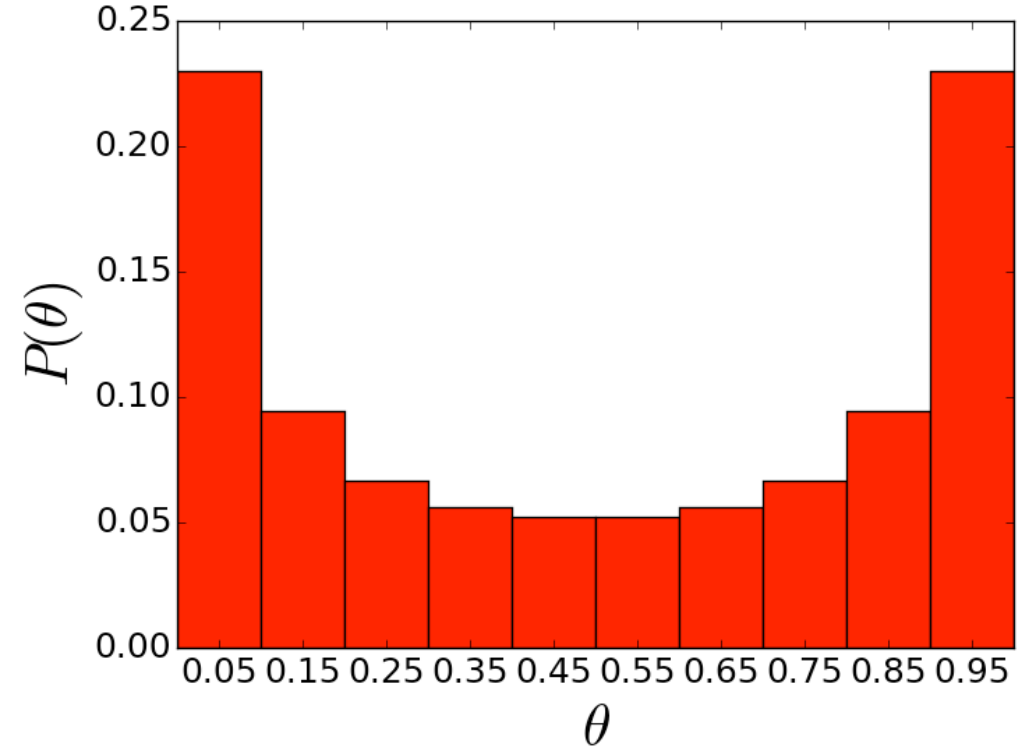
- Let’s say our learner considers 10 possible values of θ
  - 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95
- Our prior is a probability distribution: for each possible value of θ, we have to say how likely our learner thinks it is, before they have seen any data
  - High prior probability 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?

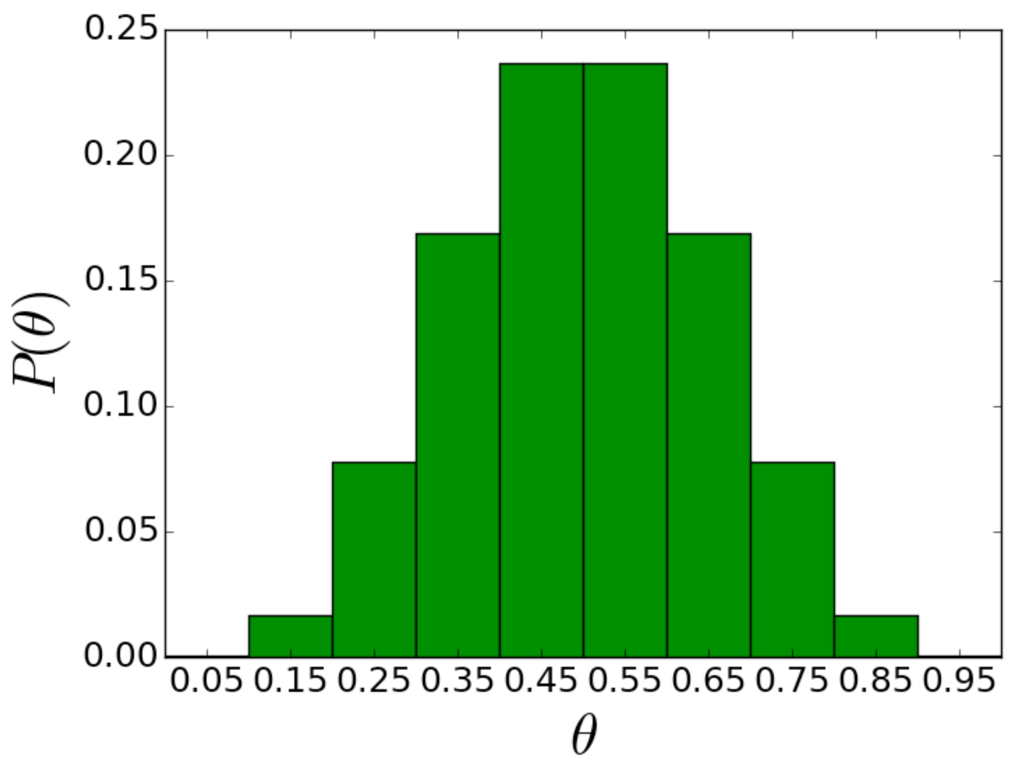
A



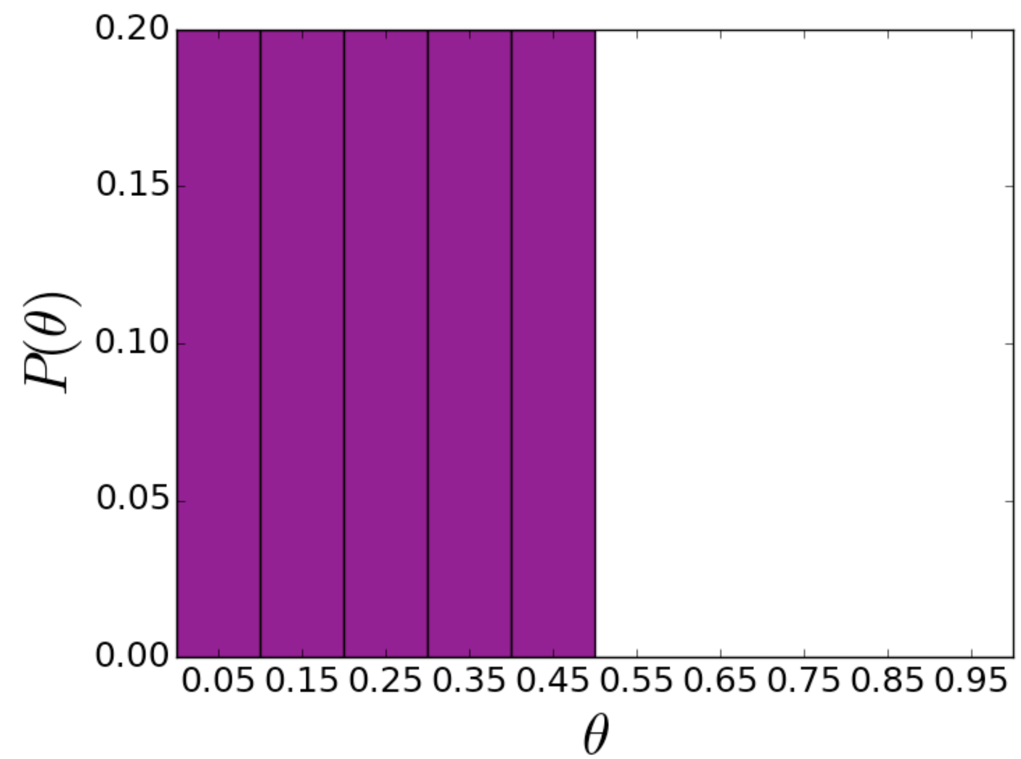
B



C

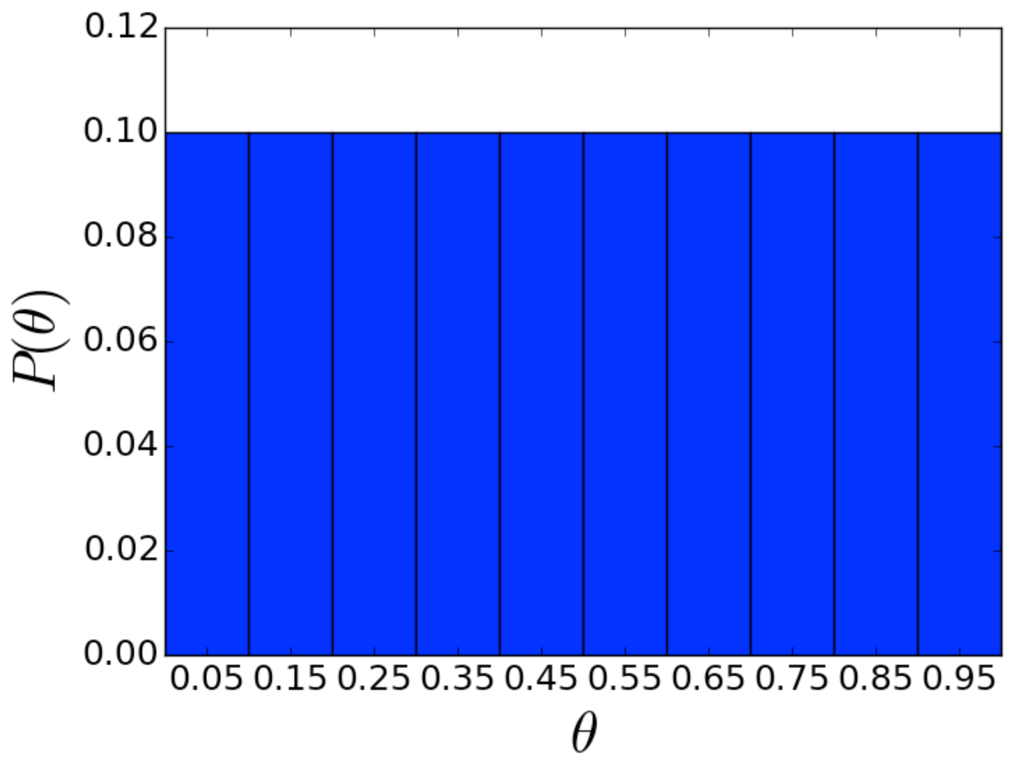


D

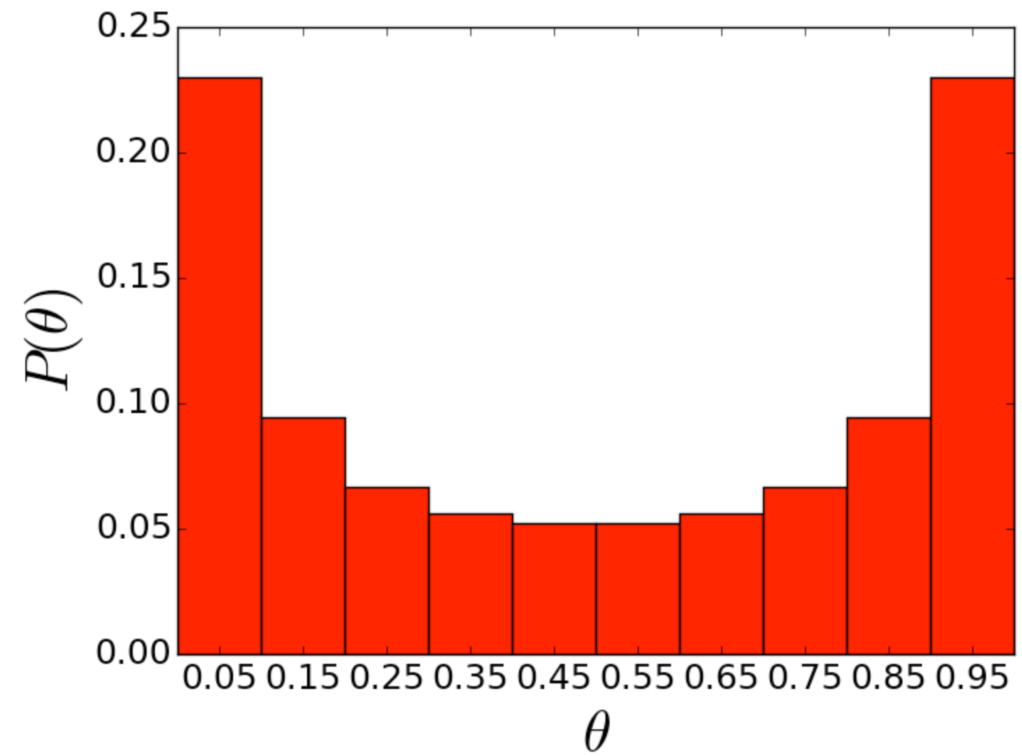


Which of these possible priors would be a good model for a **biased** learner, who thinks **each word should be used roughly equally often**?

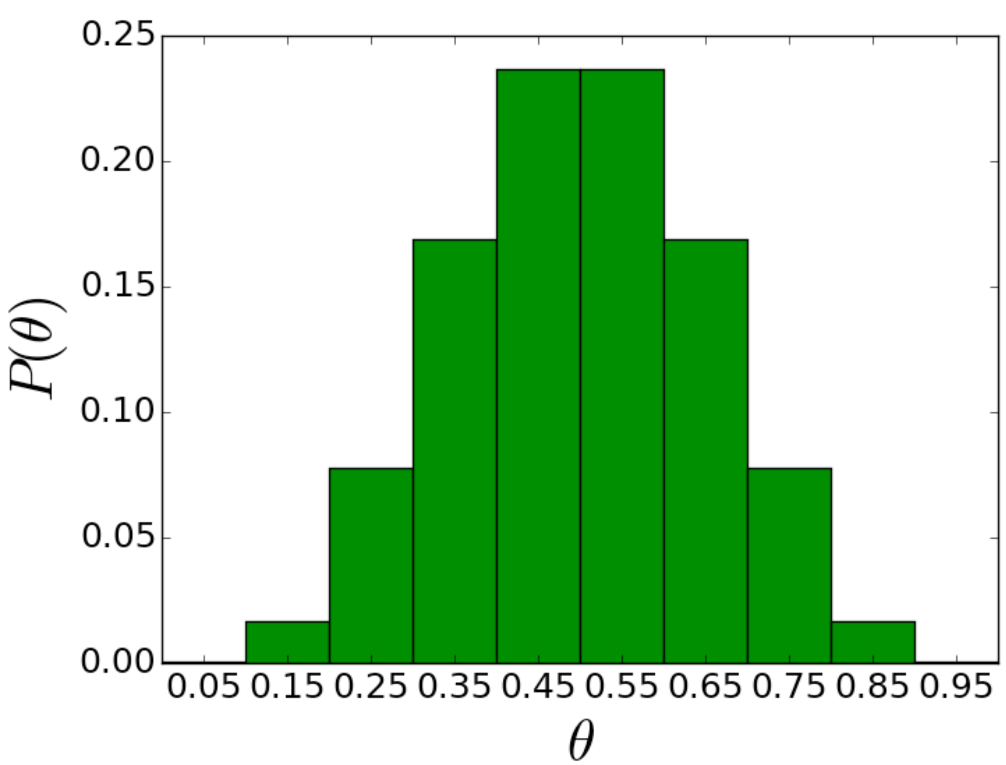
A



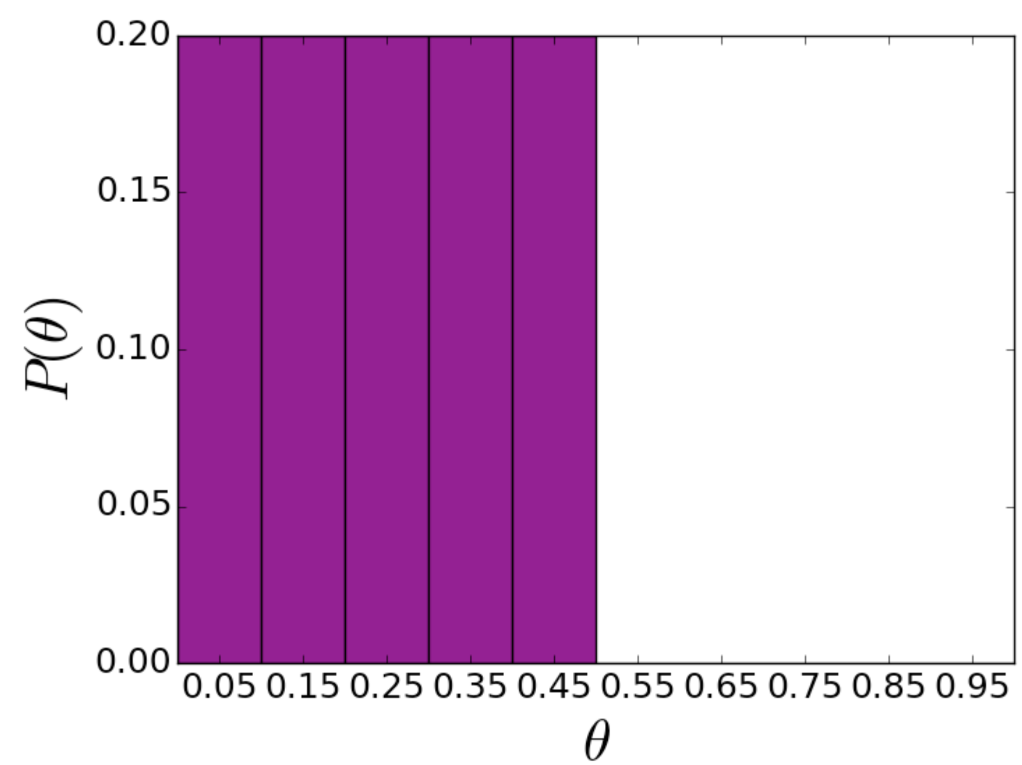
B



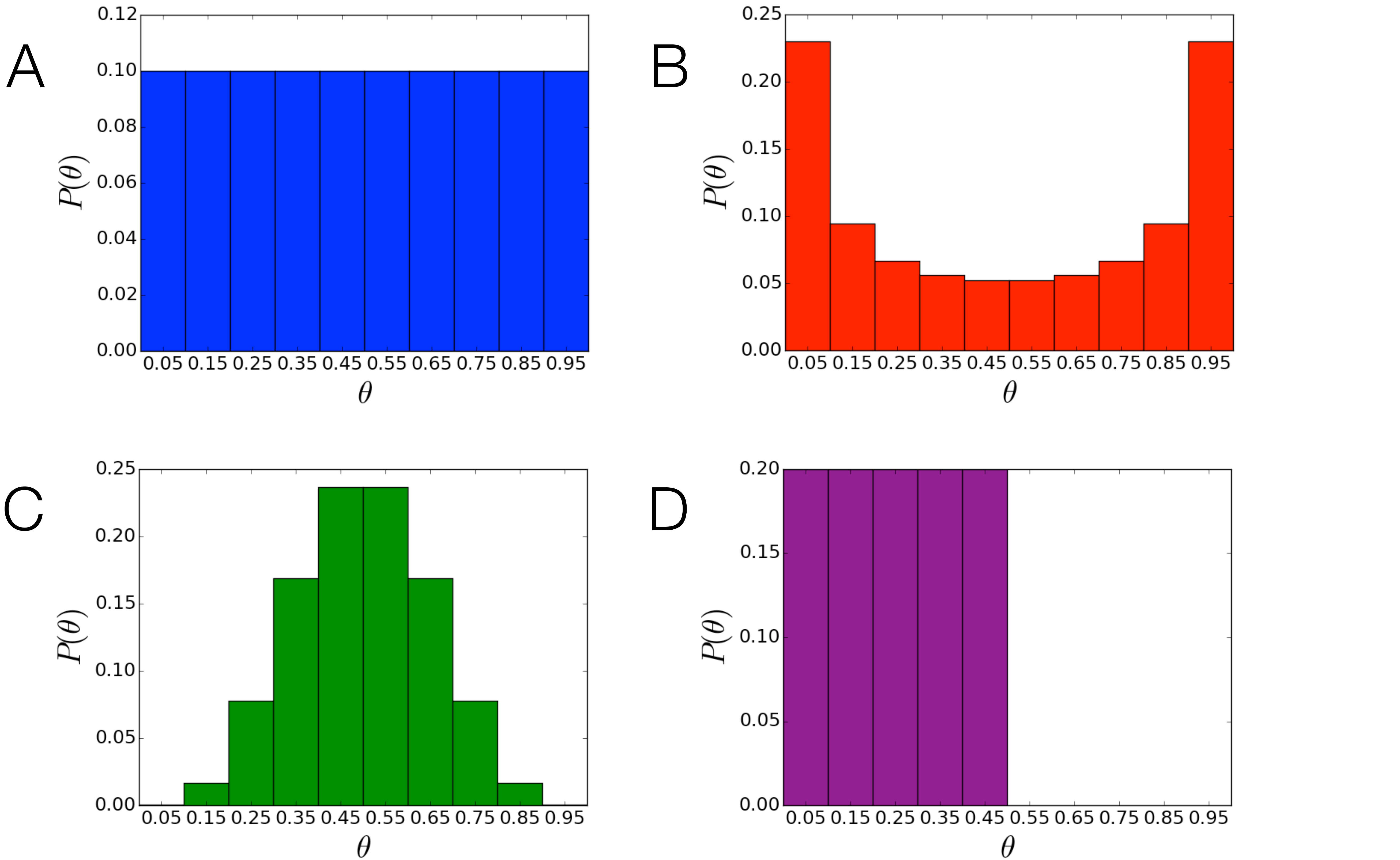
C



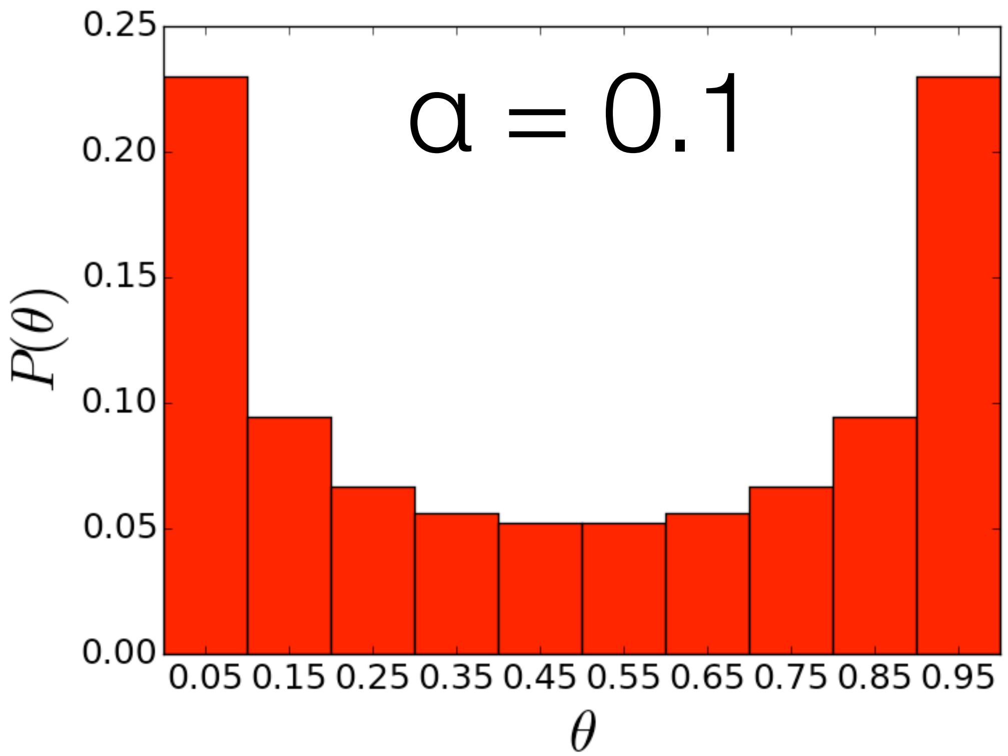
D

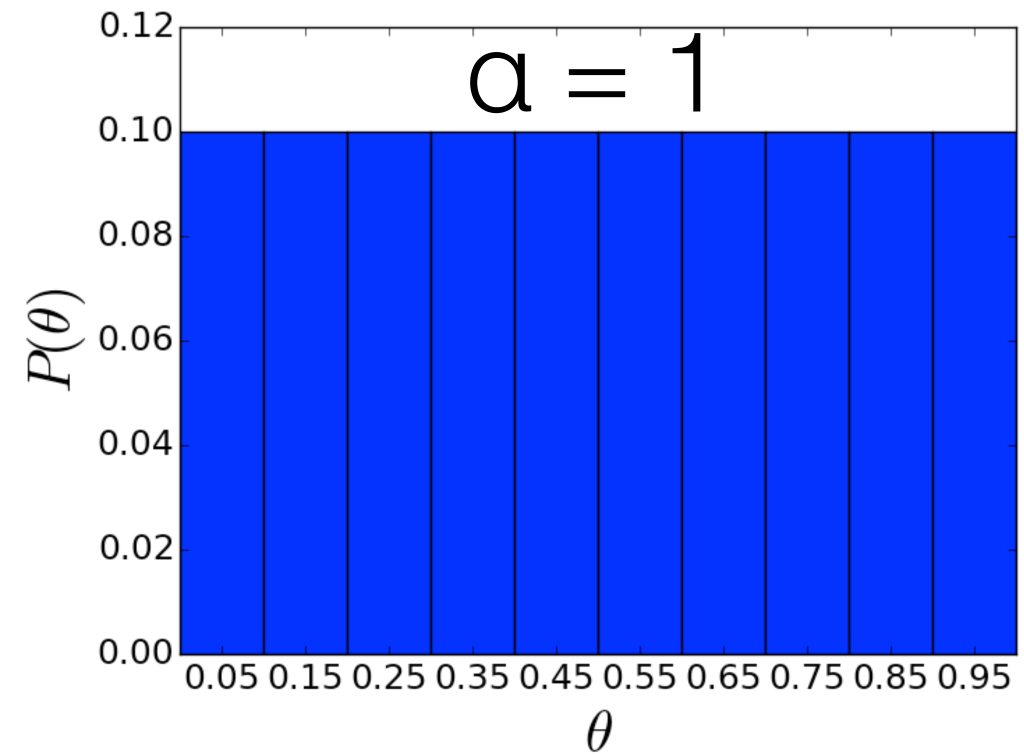


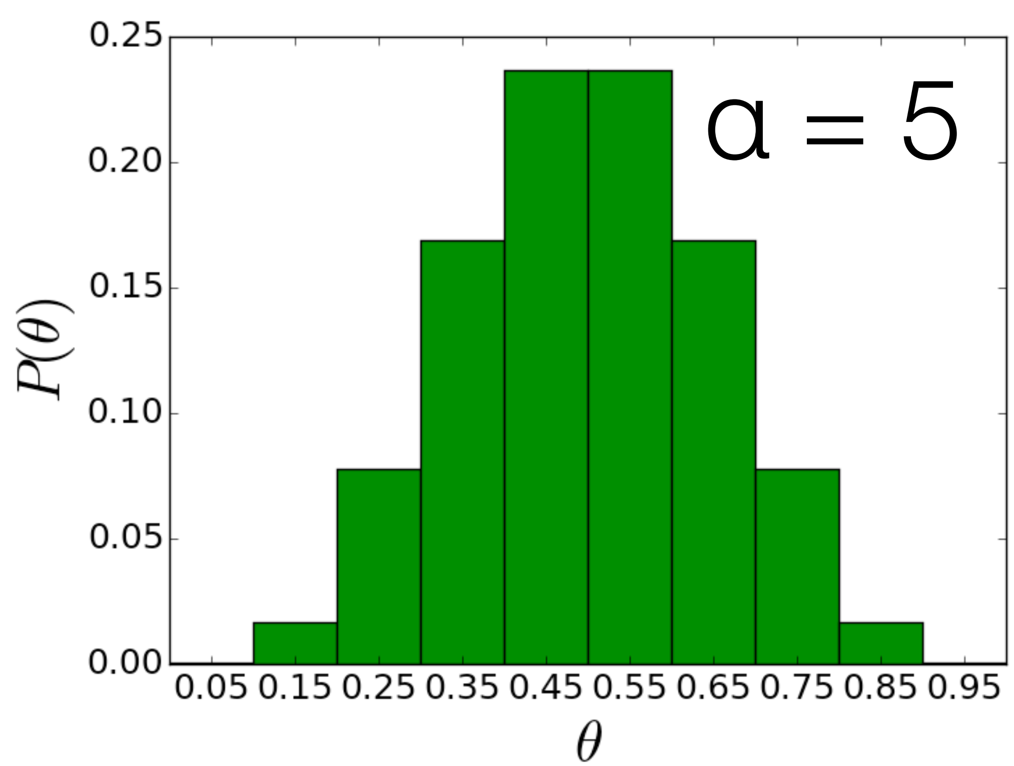
Which of these possible priors would be a good model for a **biased** learner, who thinks **only one word should be used**?



Our prior: the (symmetrical) beta distribution

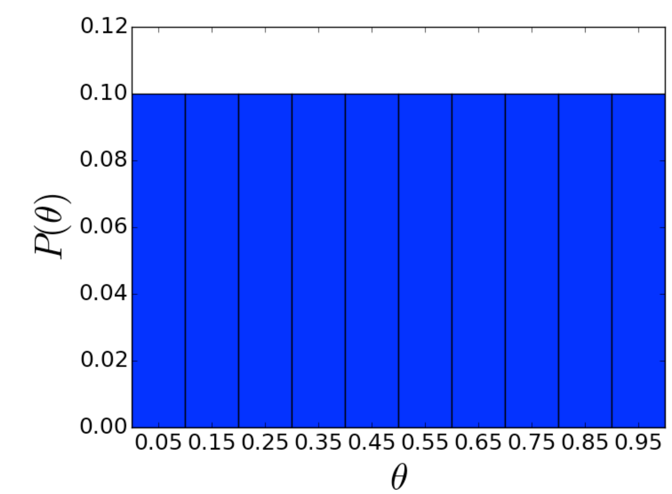






Putting it together

- 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
- 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**, and then we can just pick θ based on that (e.g. pick a value of θ according to its posterior probability)

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

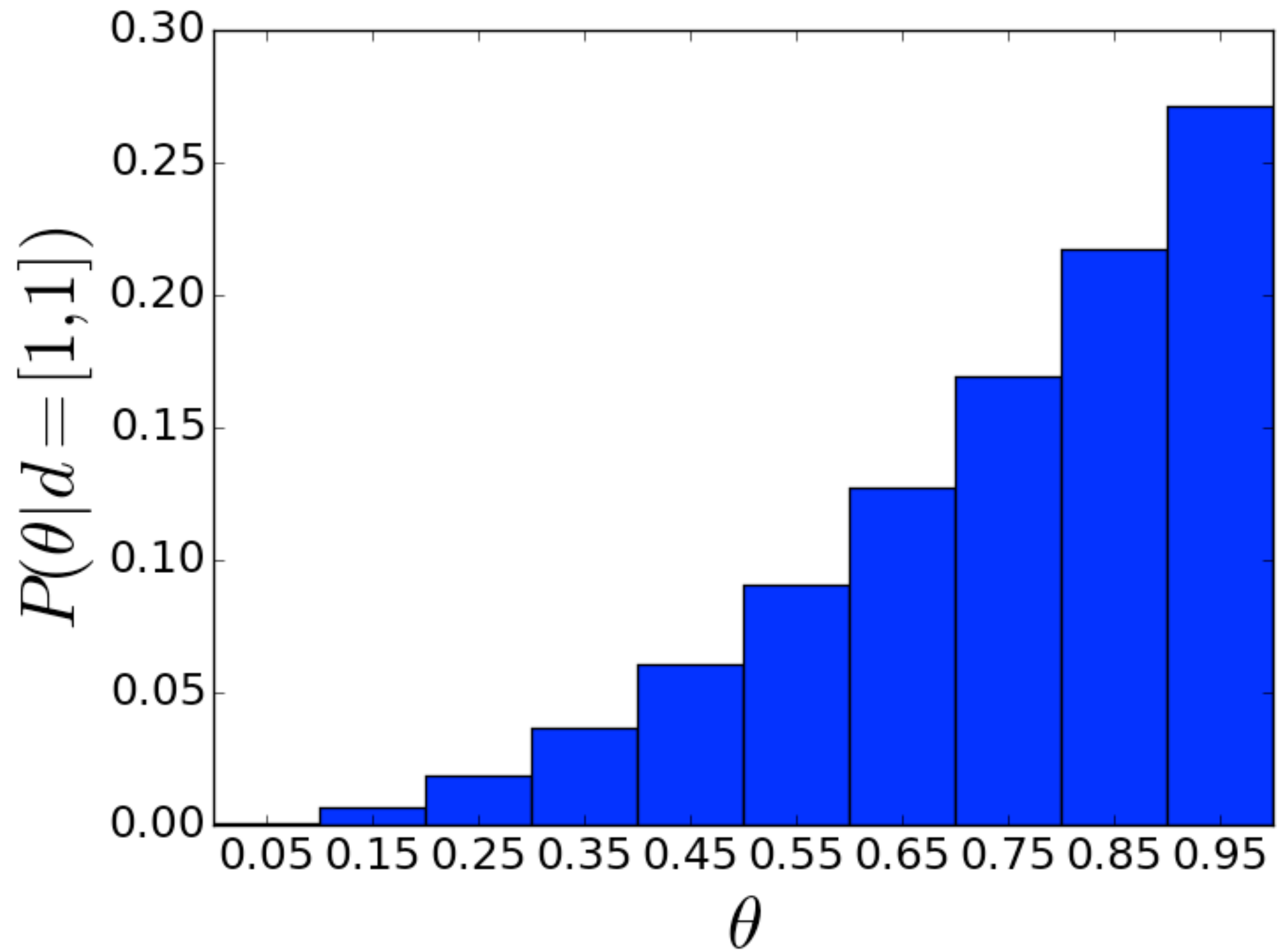
Putting it together

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

- Uniform prior, d=[1,1]
- Consider just θ=0.25 and θ=0.75. Which has higher posterior probability?
  - θ=0.75 has a higher posterior probability
- How much higher?
  - θ=0.75 is 9 times higher in posterior probability than θ=0.25

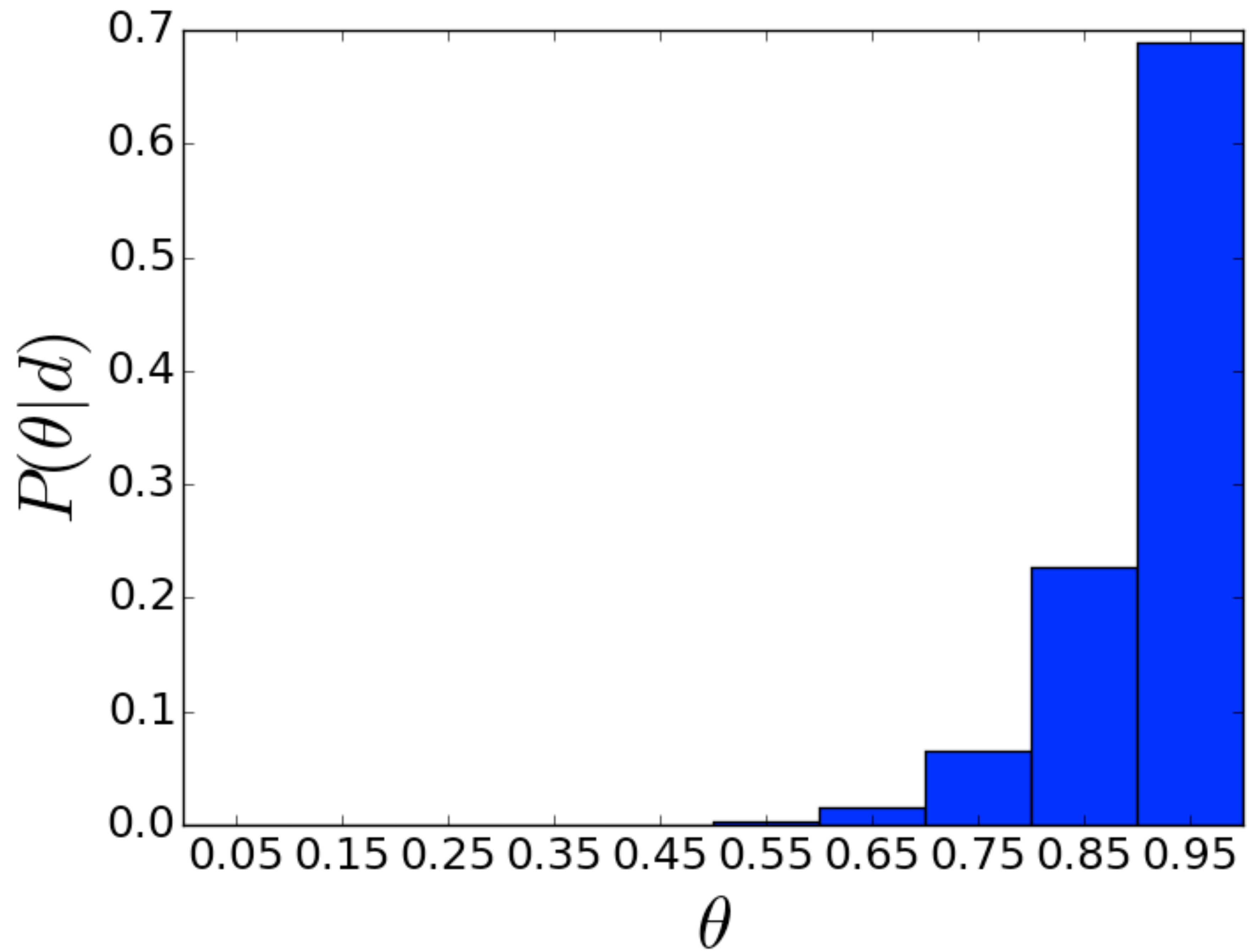
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

- Uniform prior, d=[1,1]



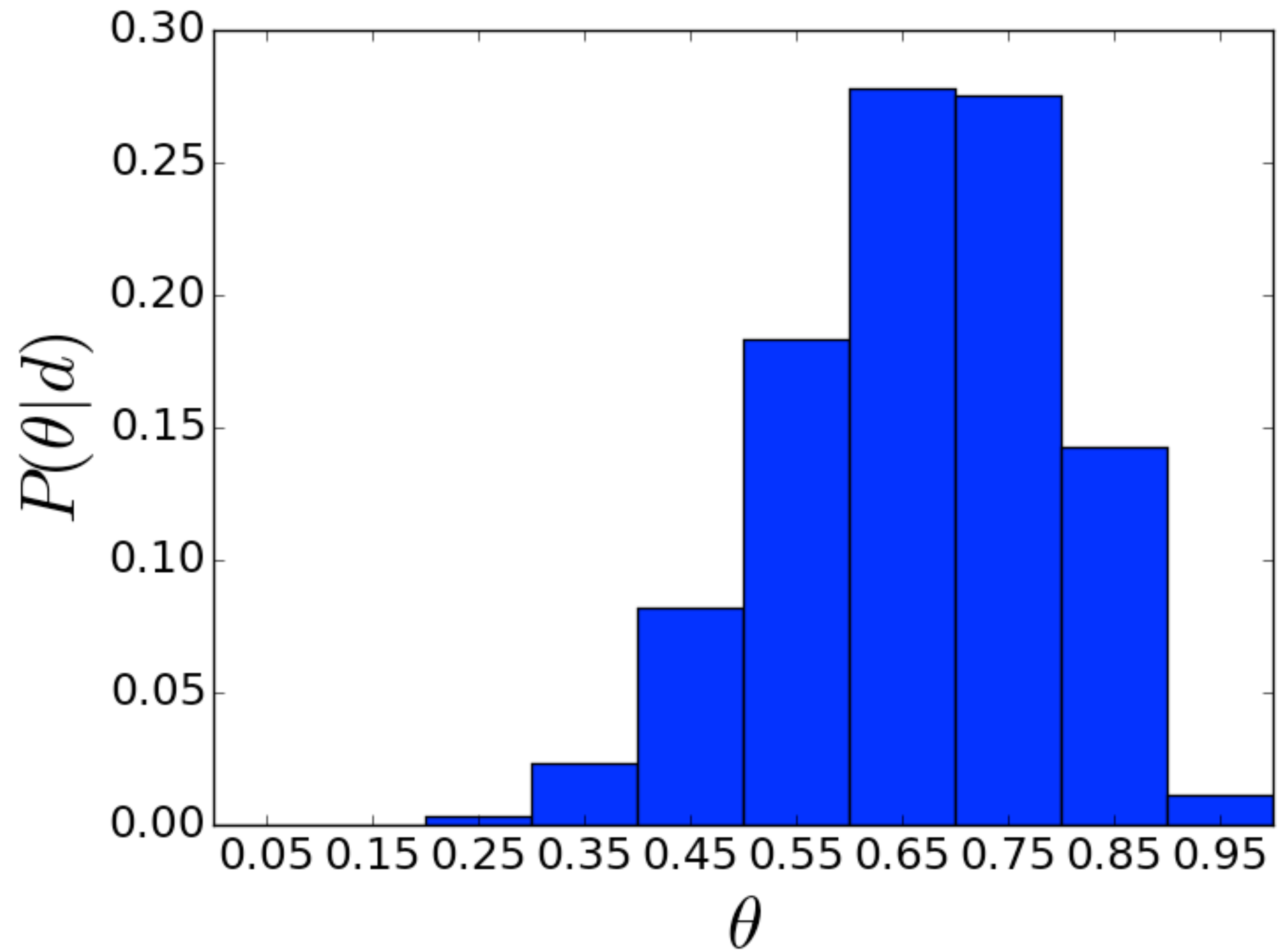
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

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



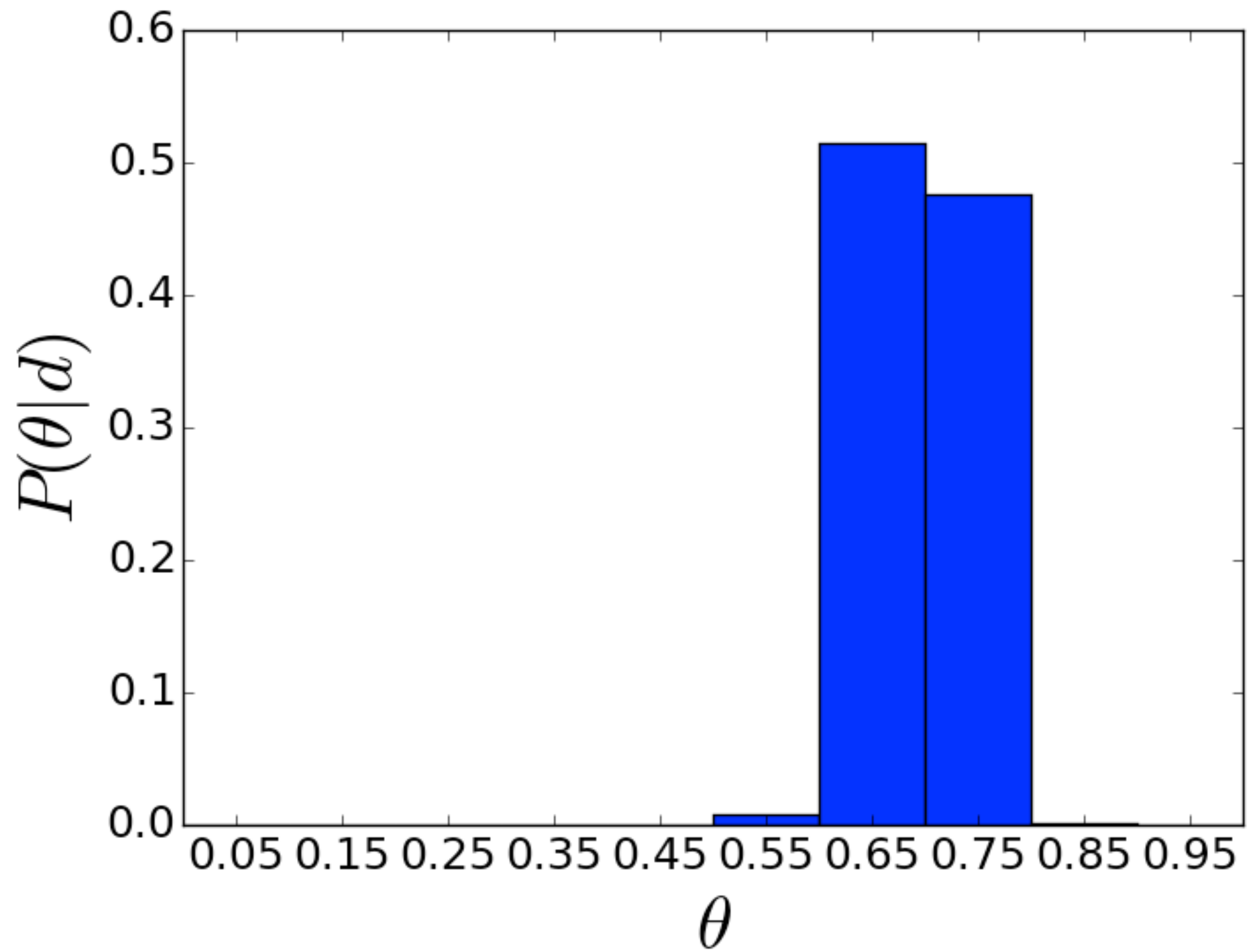
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

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



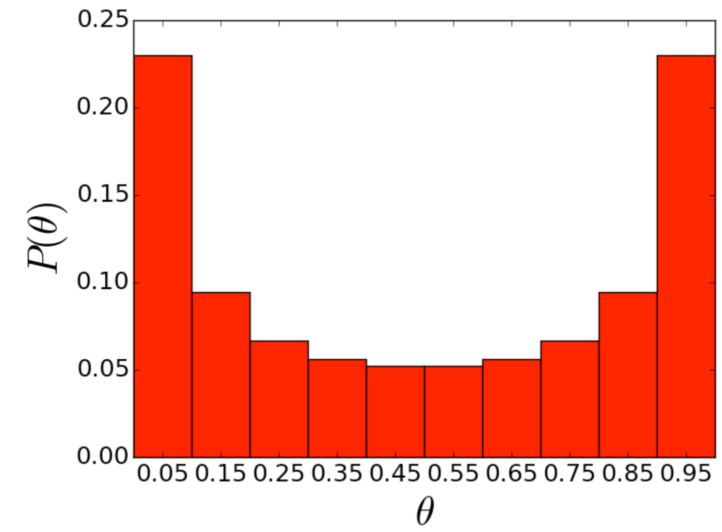
Putting it together $P(\theta|d) \propto P(d|\theta)P(\theta)$

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



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

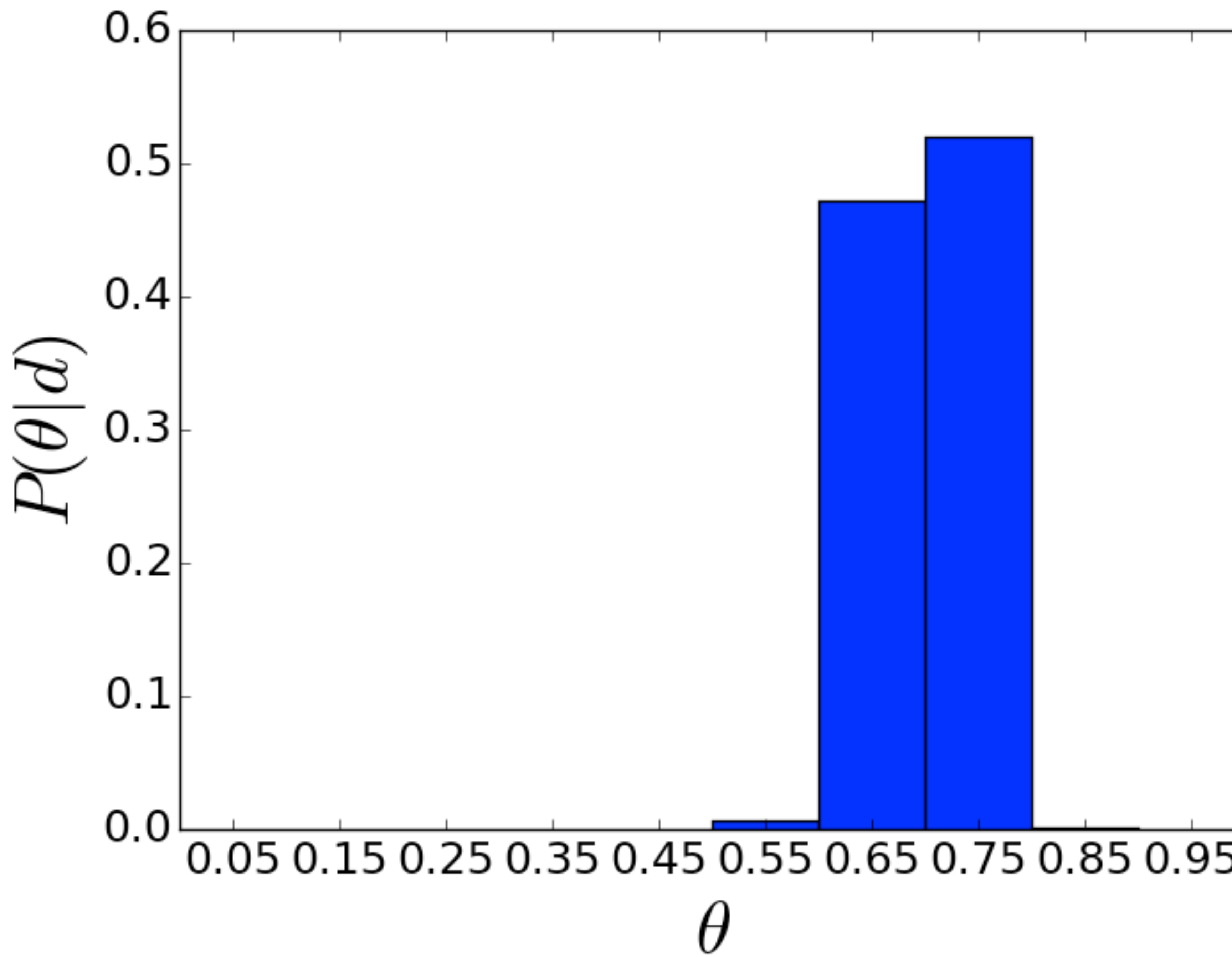
- Regularity prior, d= [70 occurrences of word 1, 30 of word 0]



Putting it together

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

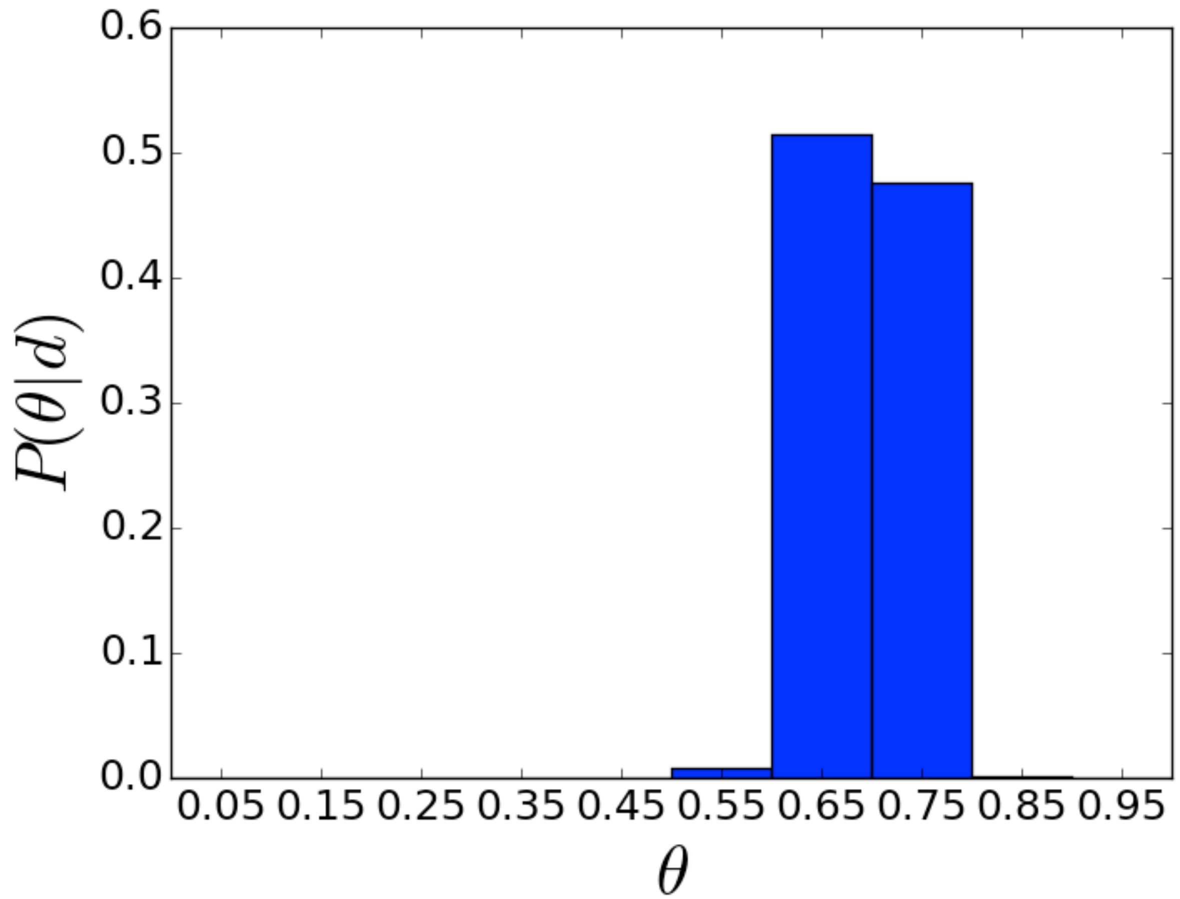
- Regularity prior, d= [70 occurrences of word 1, 30 of word 0]



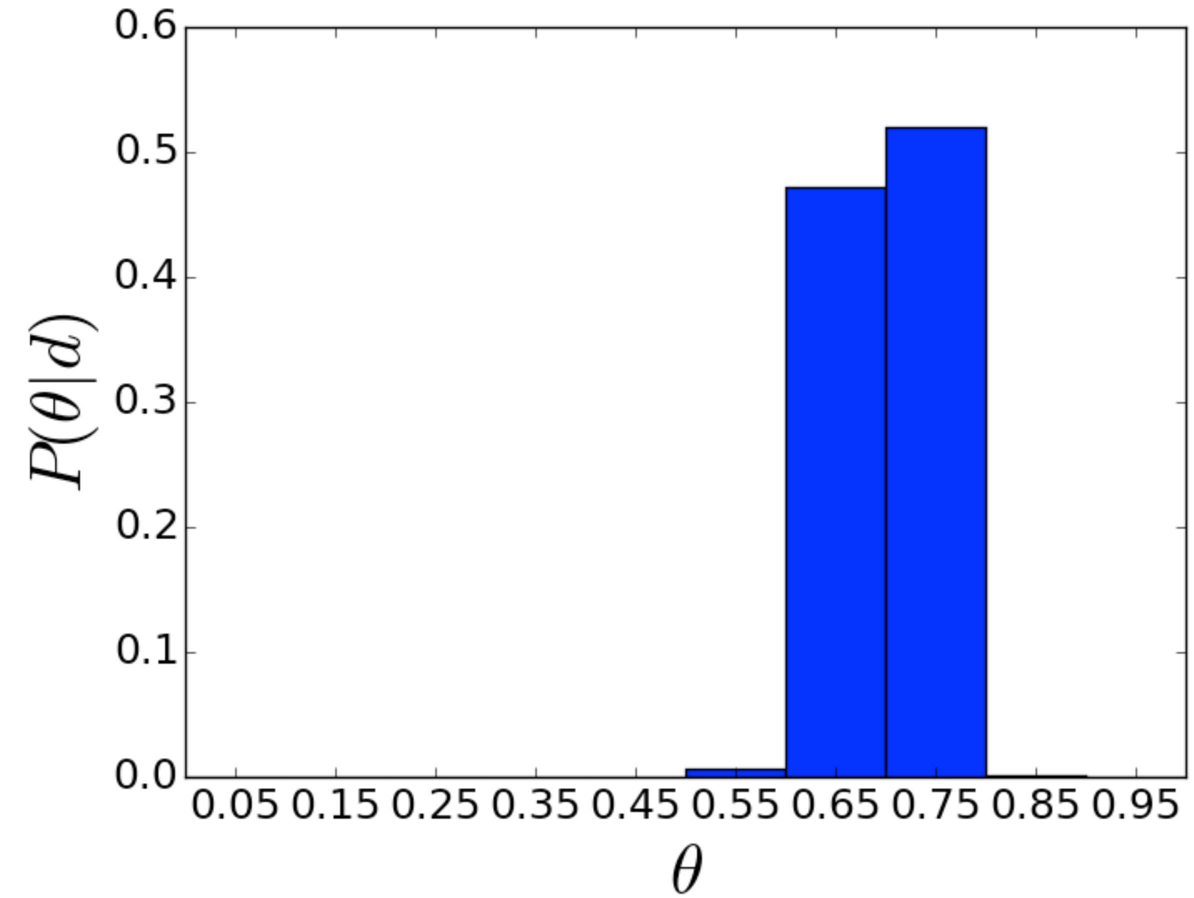
Data obscures the prior

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

Unbiased learner

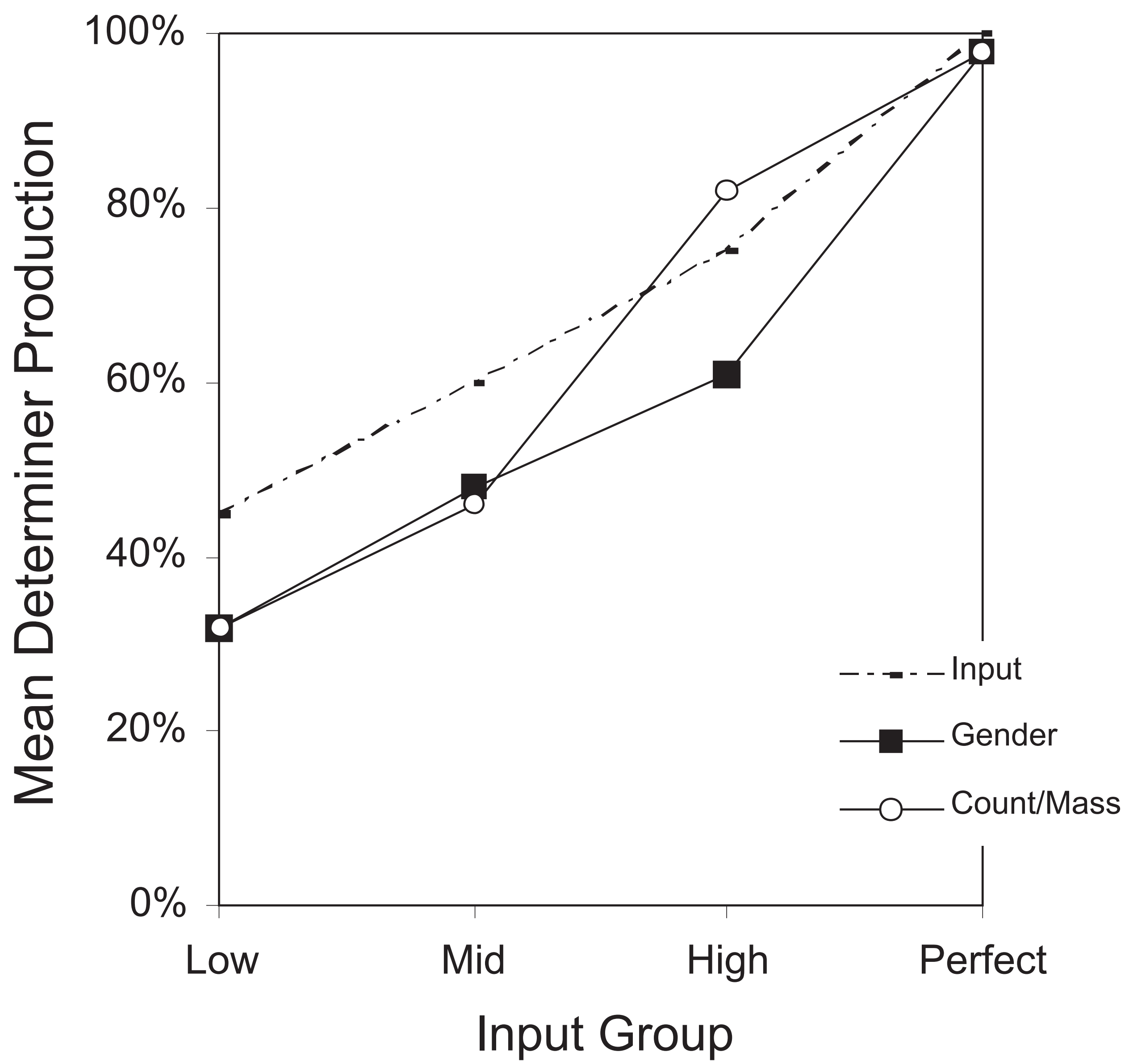


Biased learner

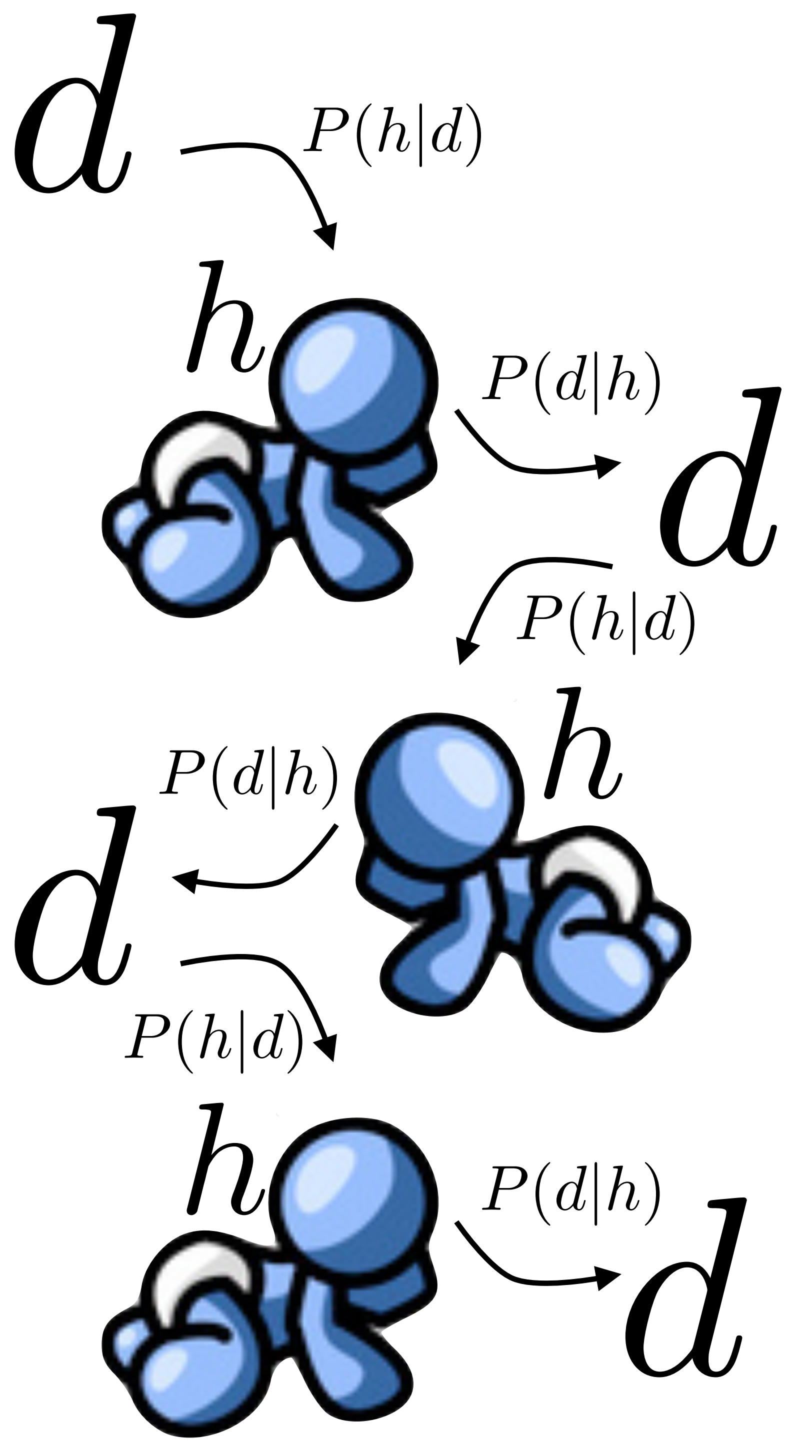


Data obscures the prior $P(\theta|d) \propto P(d|\theta)P(\theta)$

Unbiased learner? Biased learner?

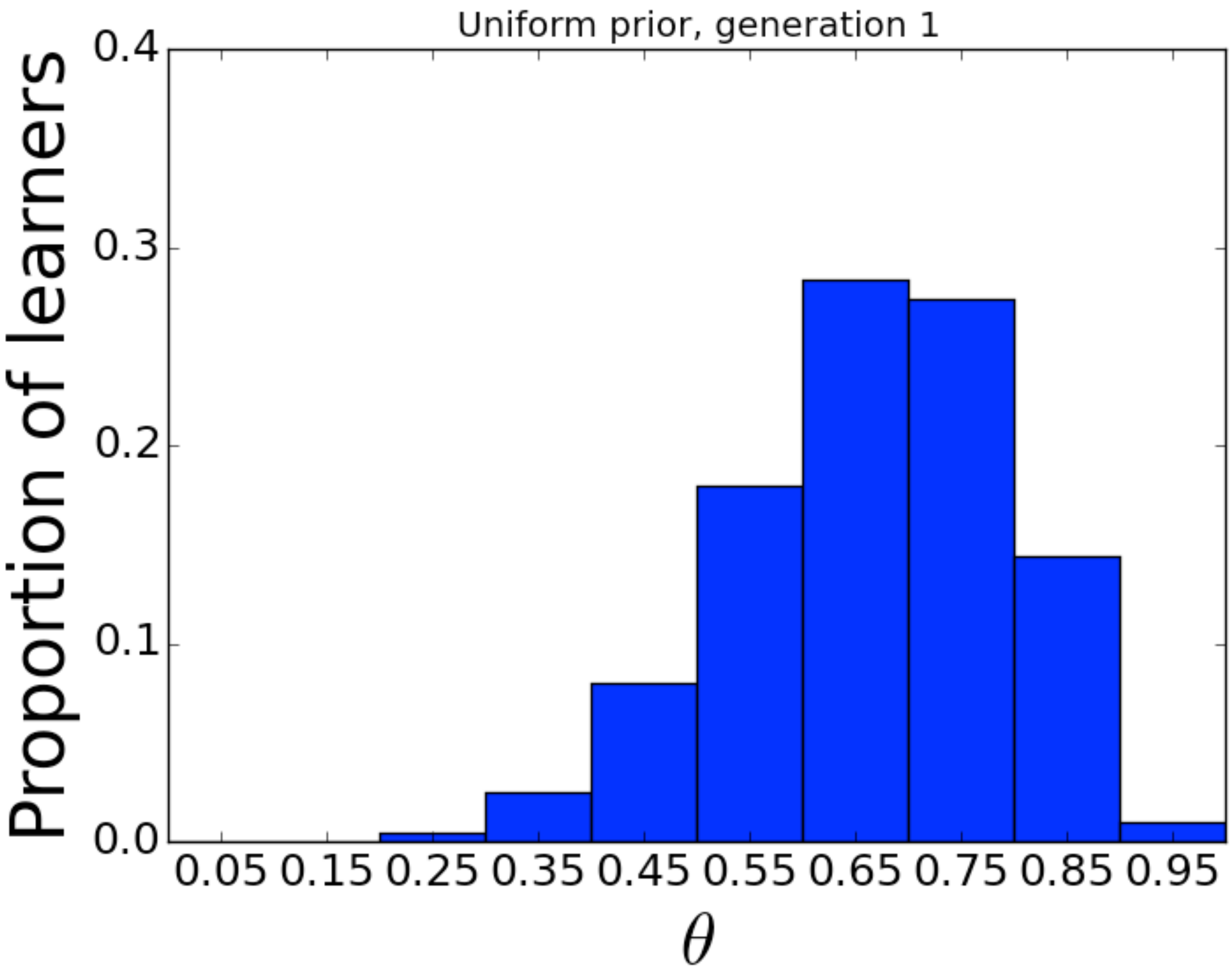


The solution: iterated learning

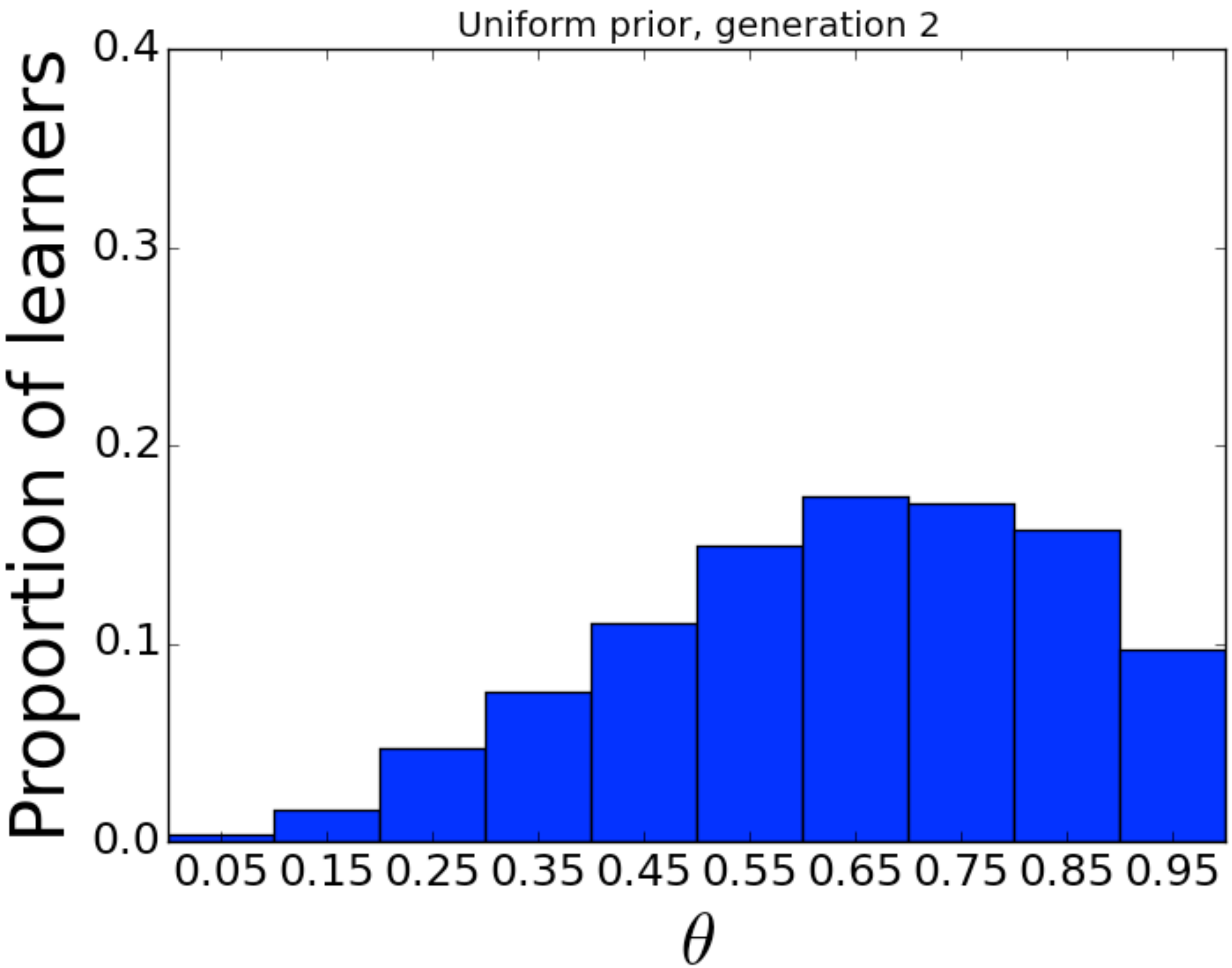


Over time, the bias
will reveal itself?

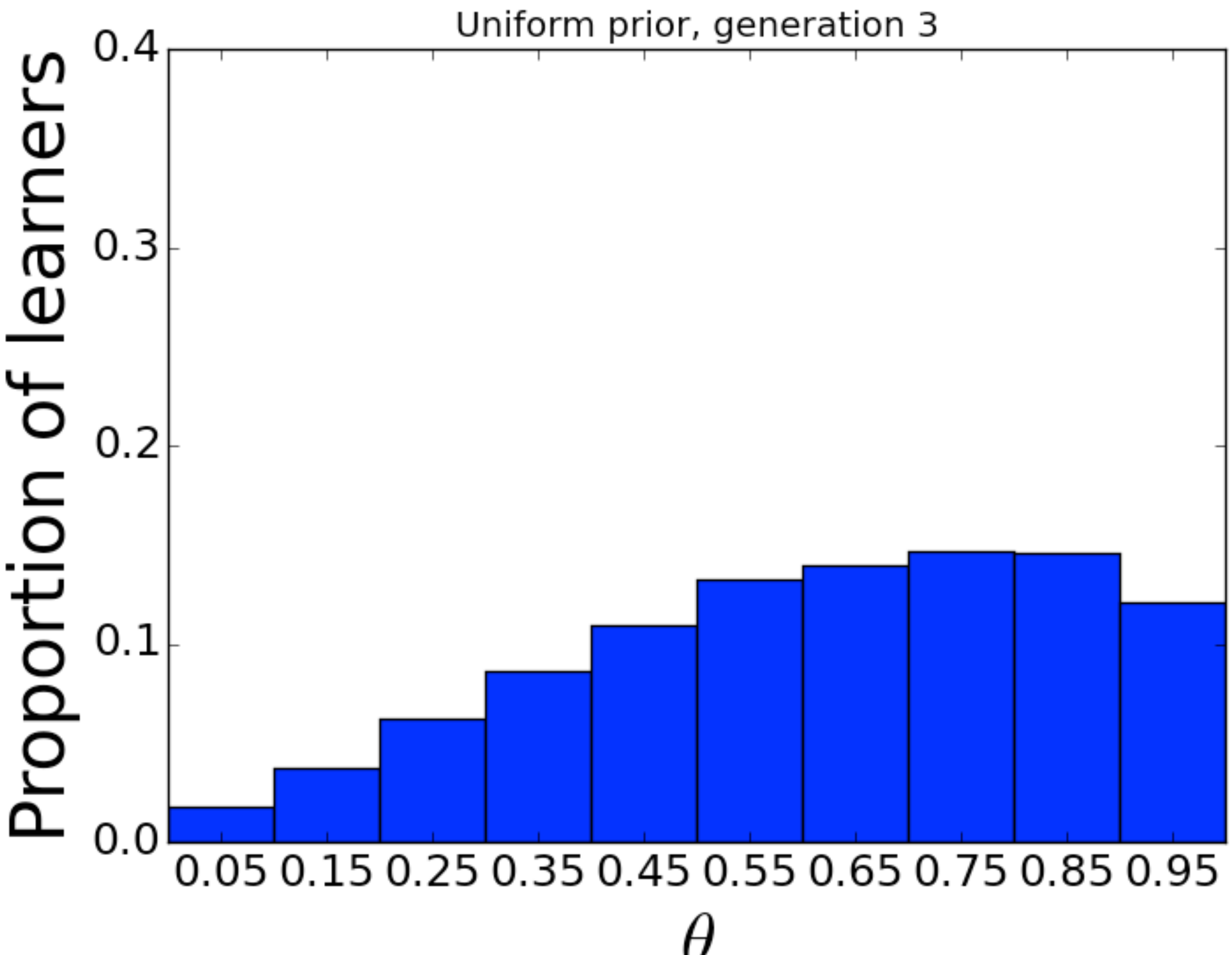
Watching the prior reveal itself



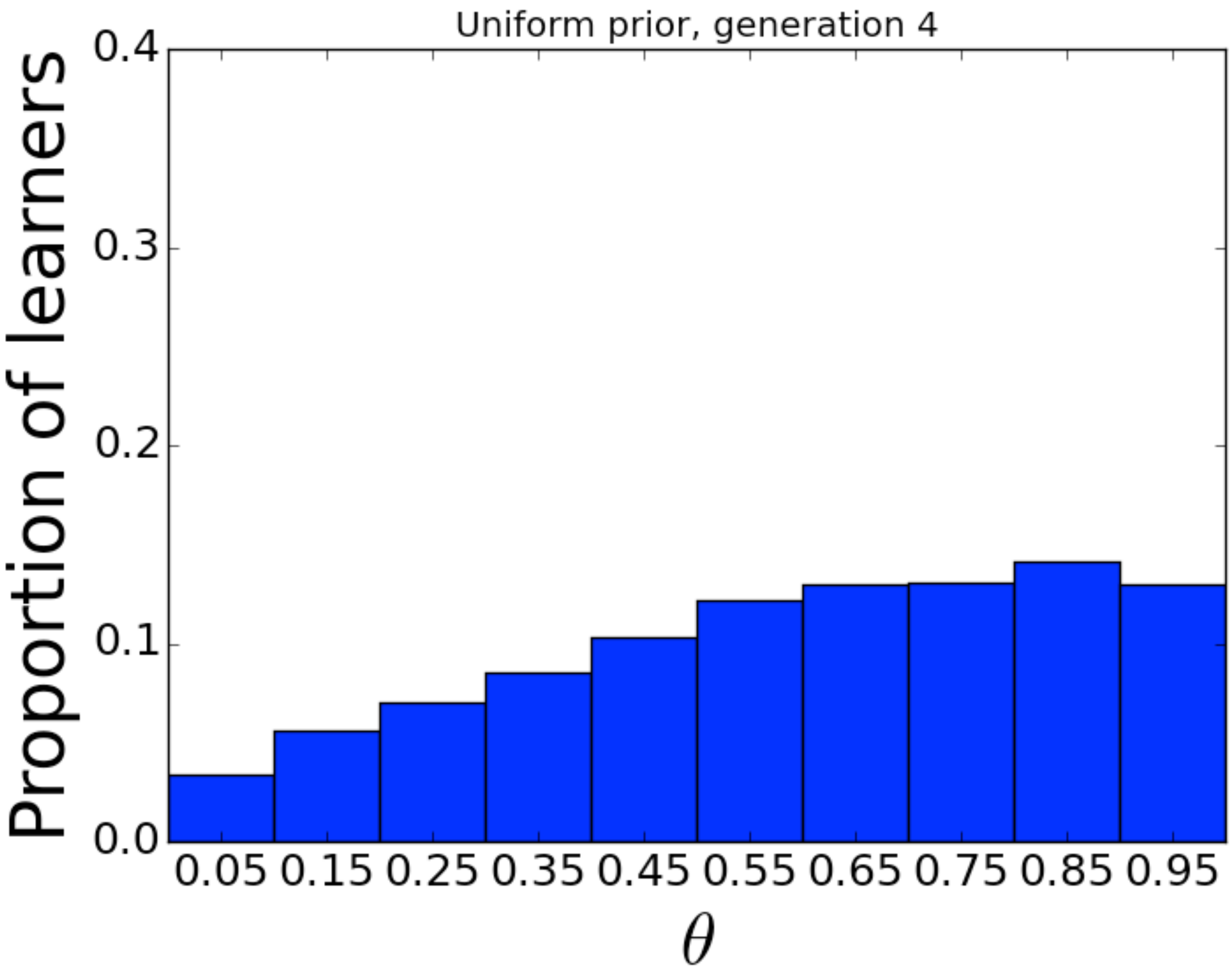
Watching the prior reveal itself



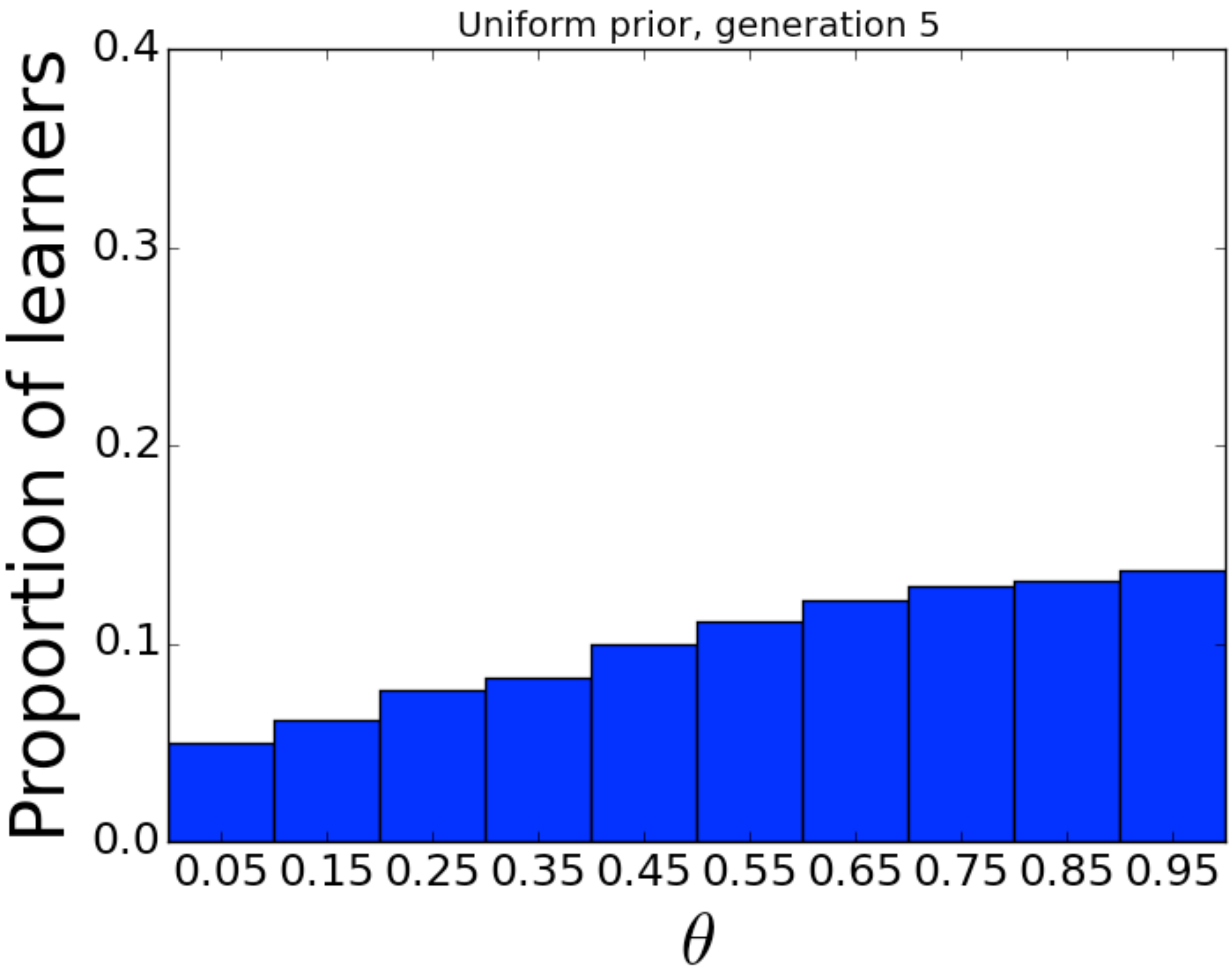
Watching the prior reveal itself



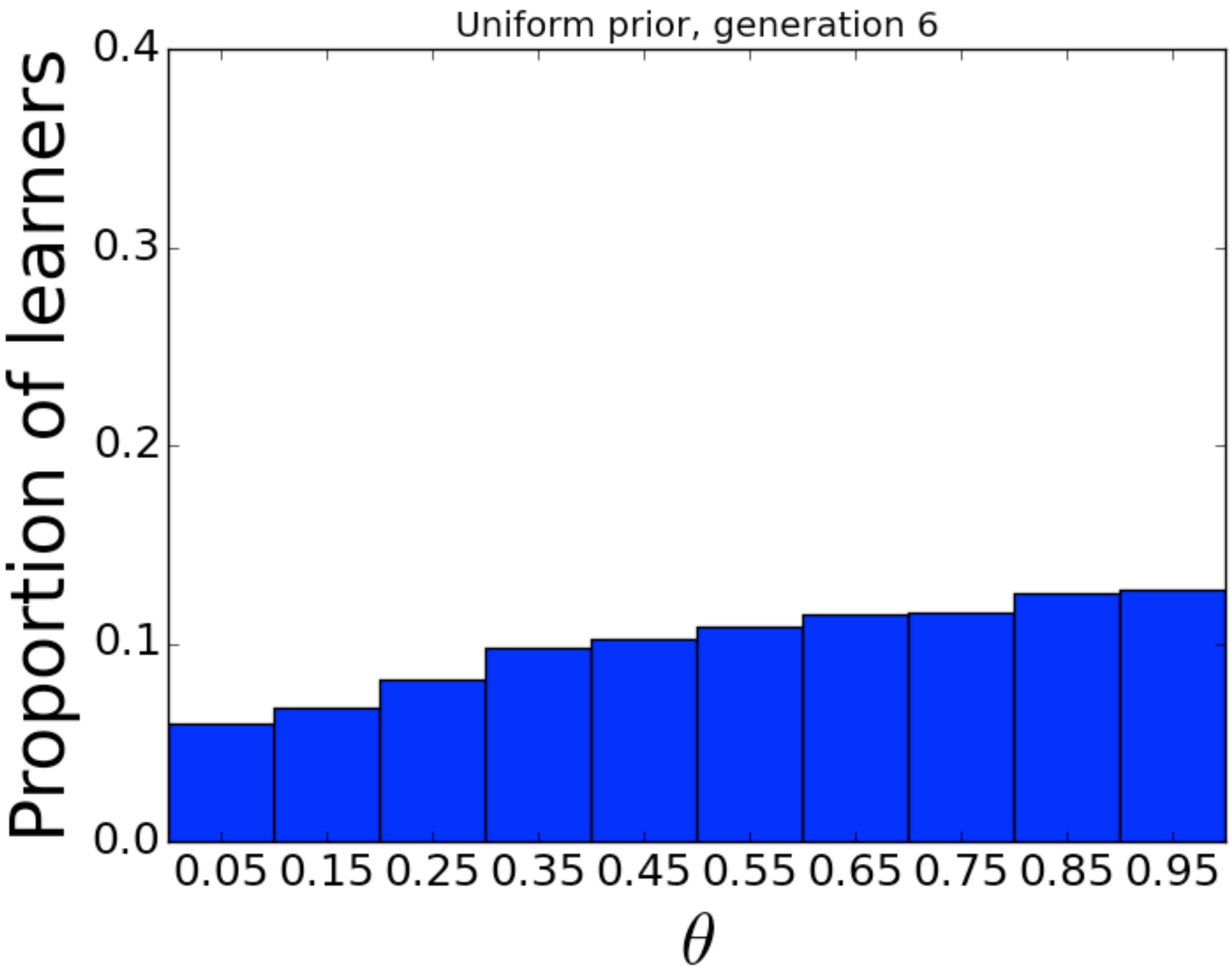
Watching the prior reveal itself



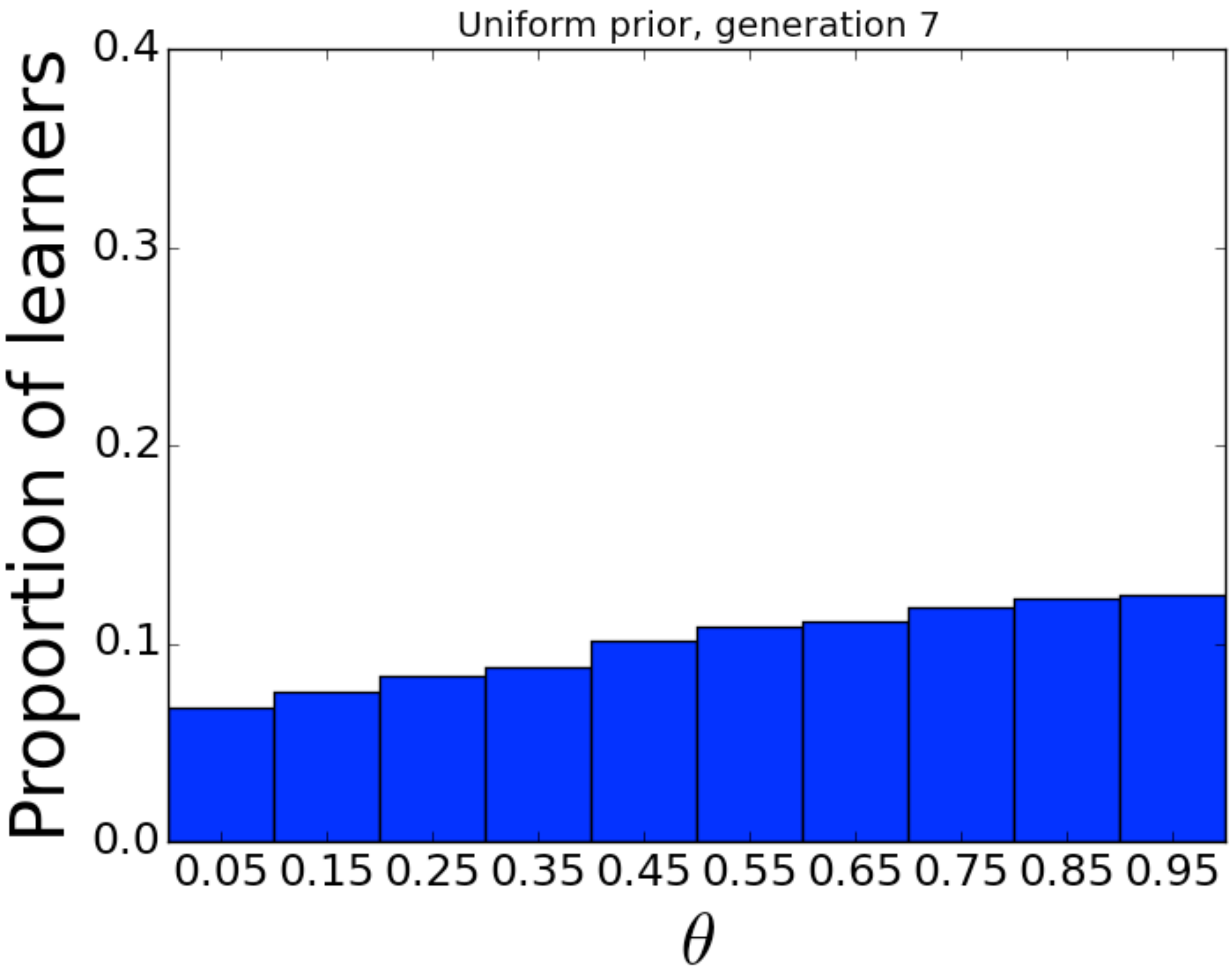
Watching the prior reveal itself



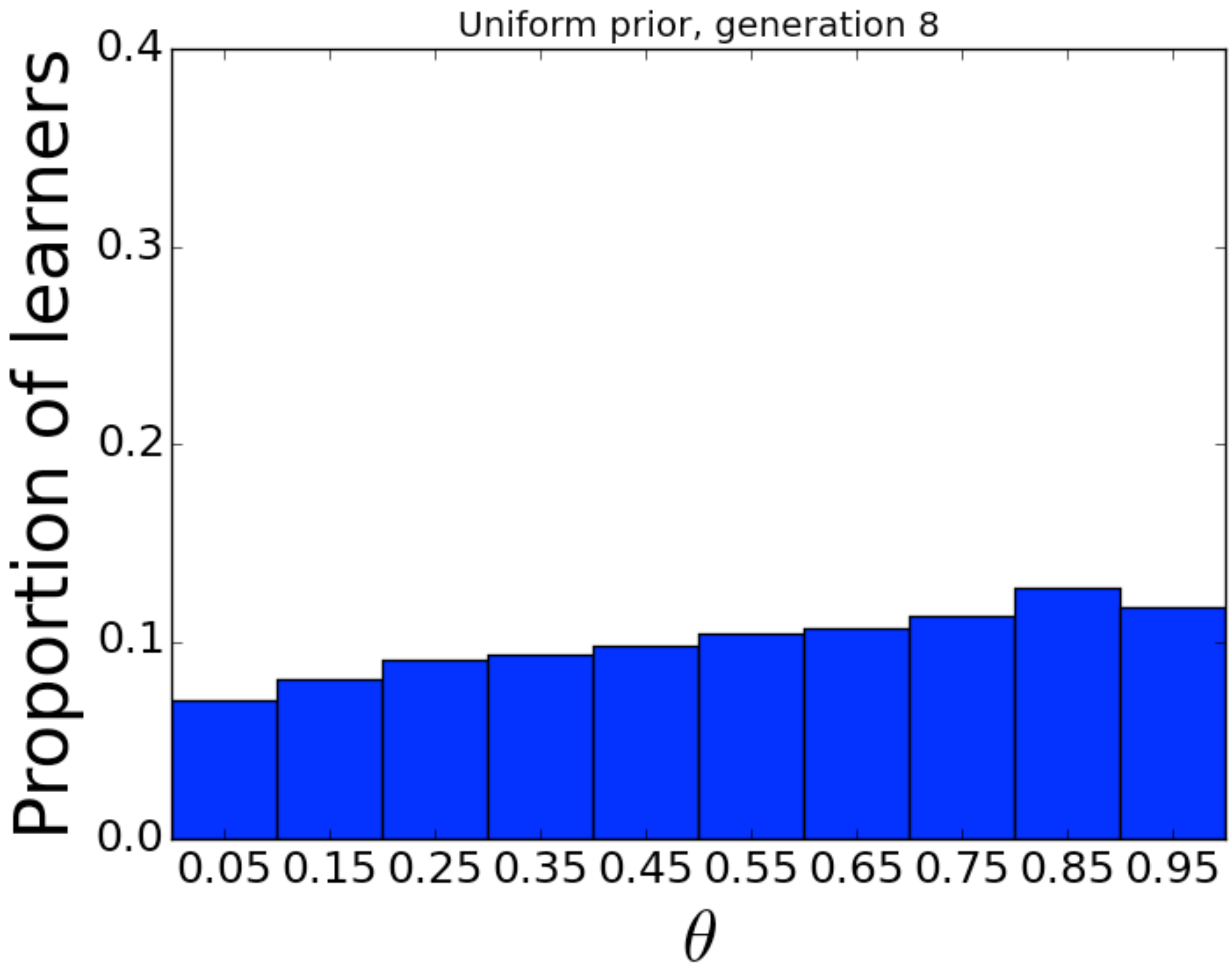
Watching the prior reveal itself



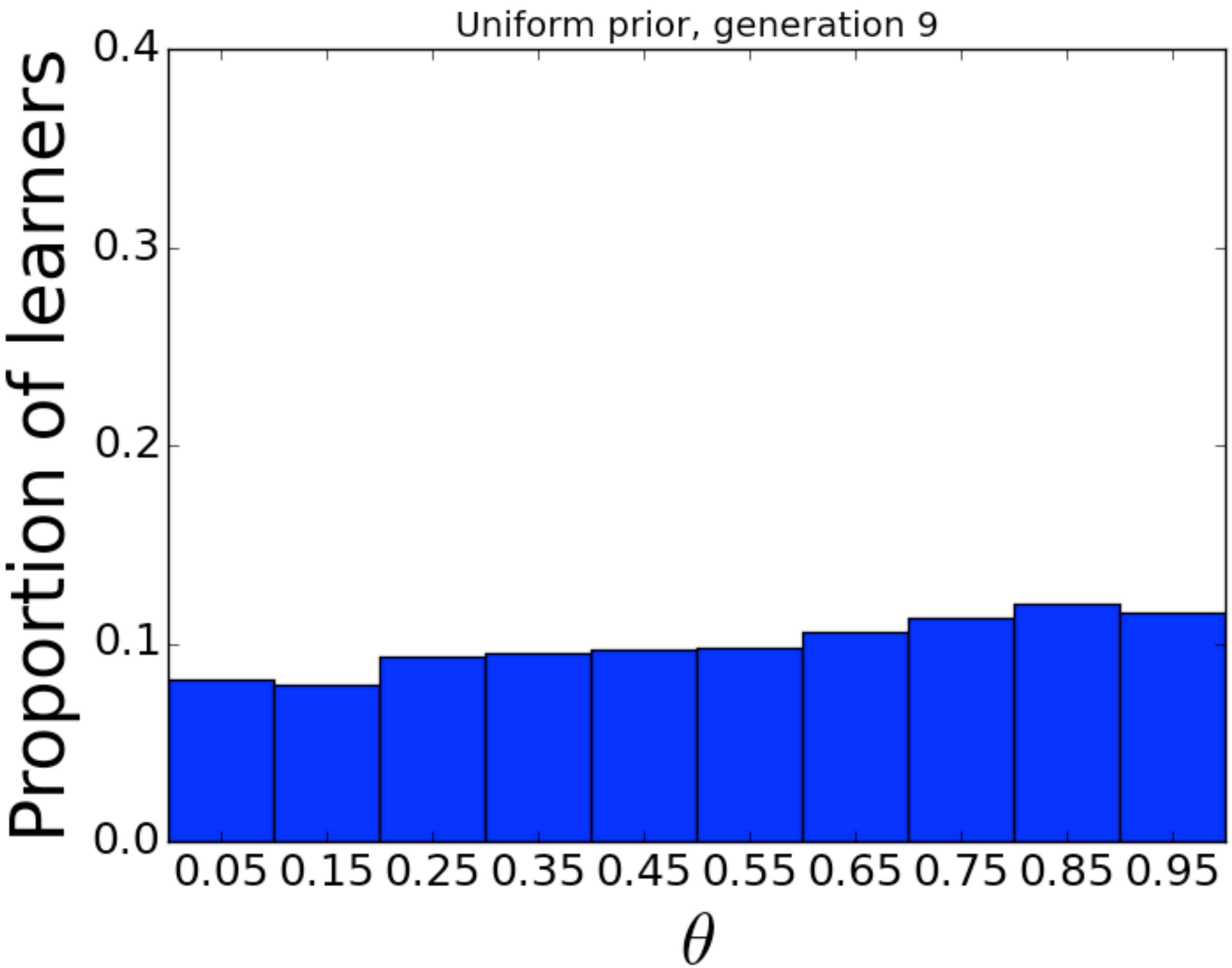
Watching the prior reveal itself



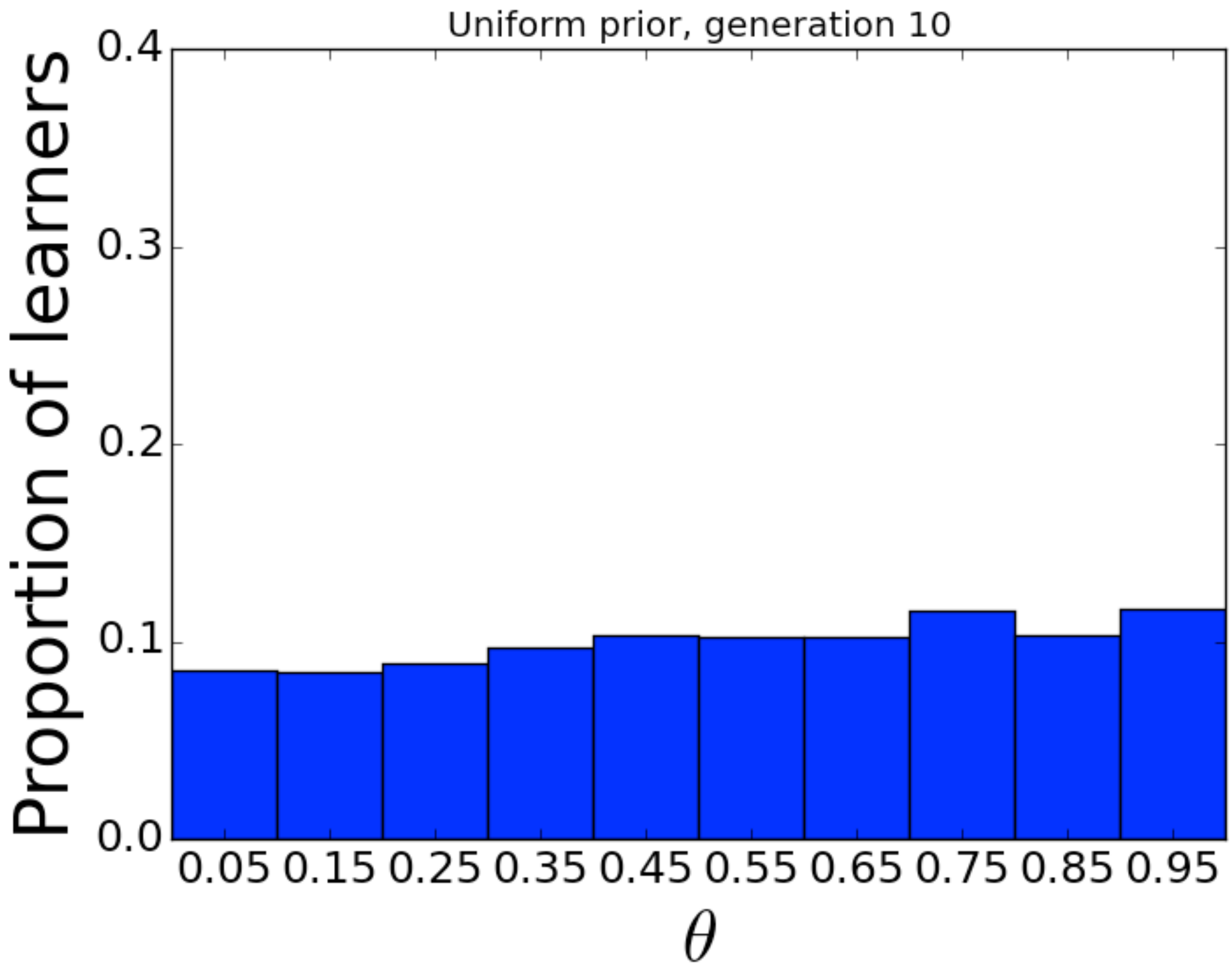
Watching the prior reveal itself



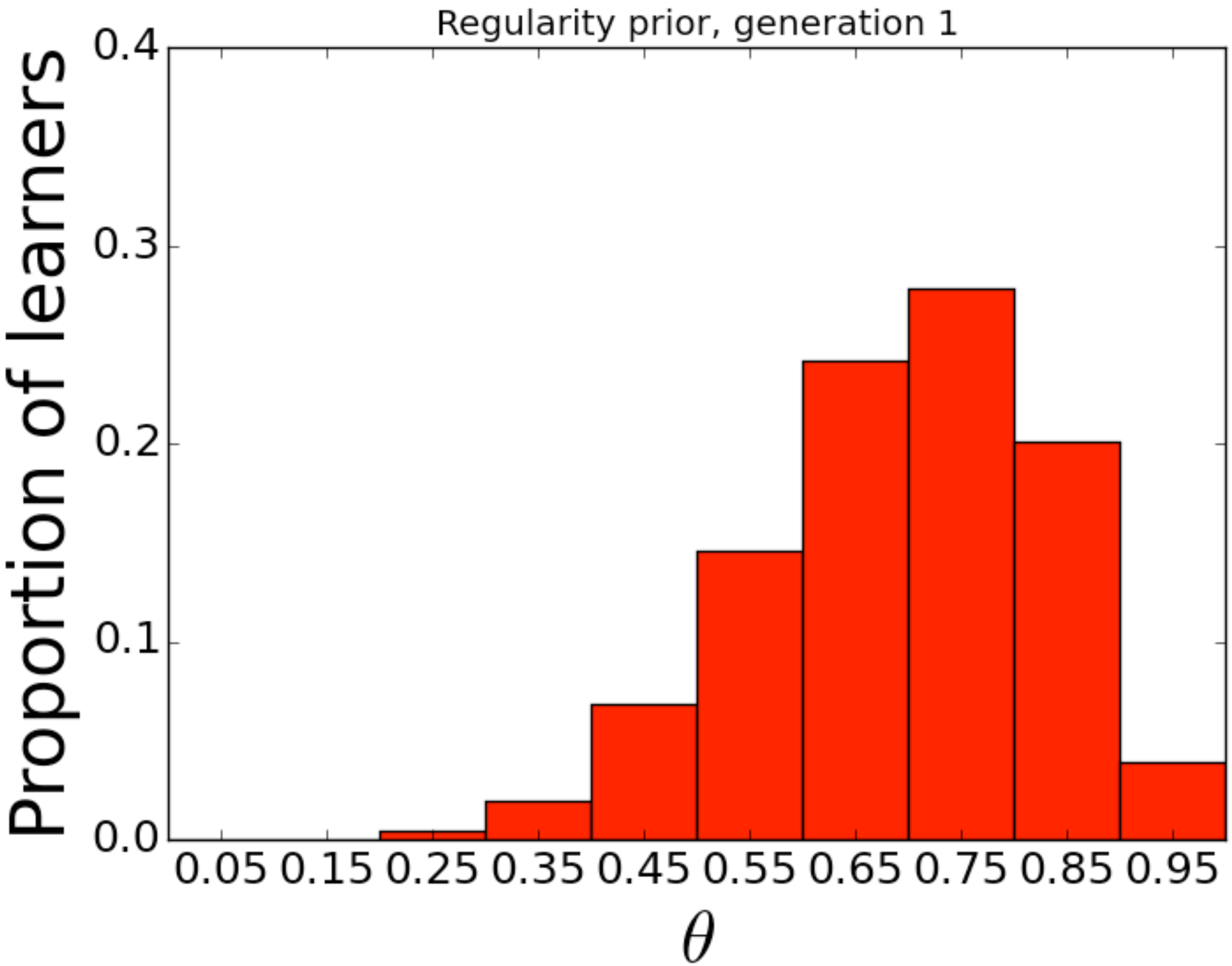
Watching the prior reveal itself



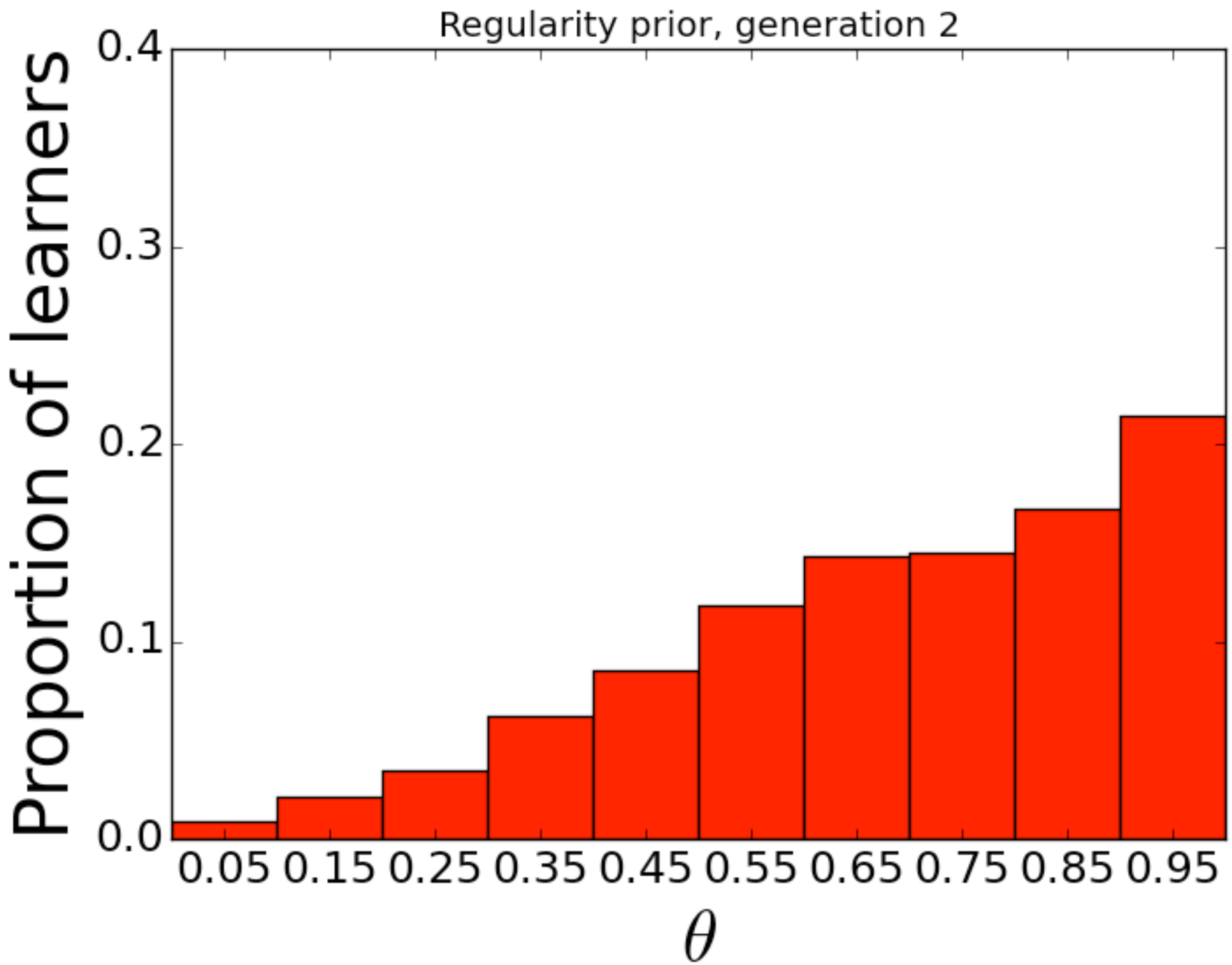
Watching the prior reveal itself



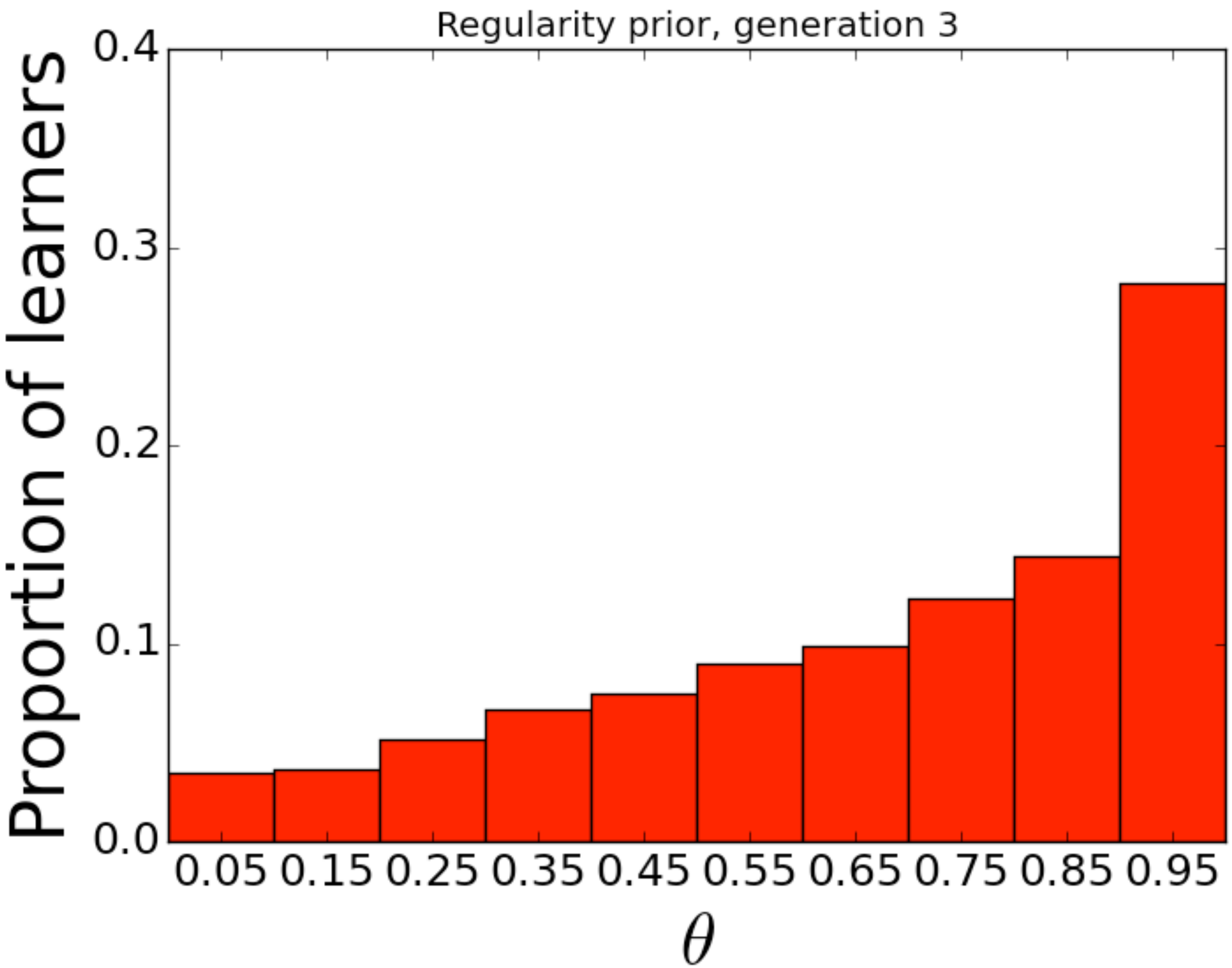
Watching the prior reveal itself



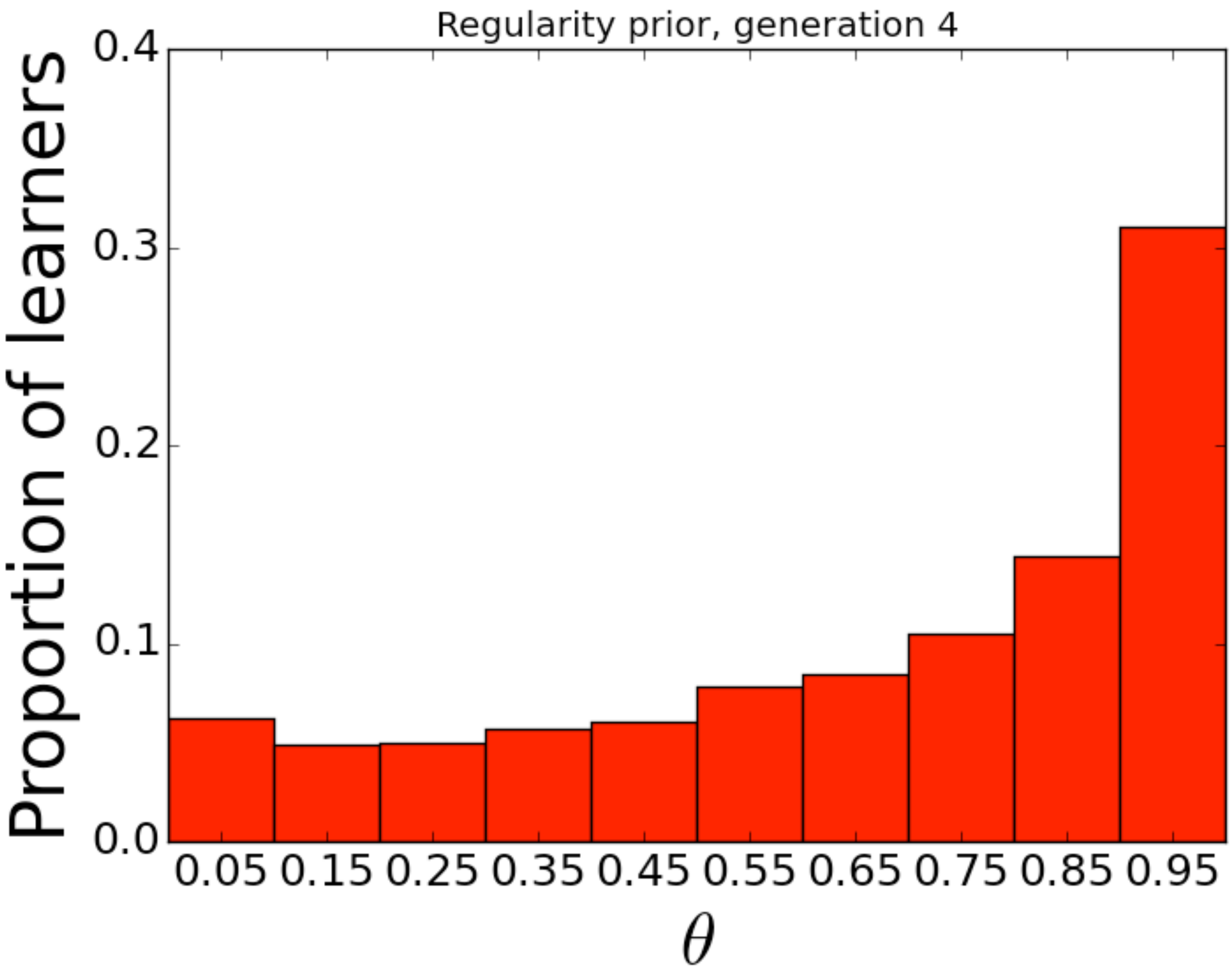
Watching the prior reveal itself



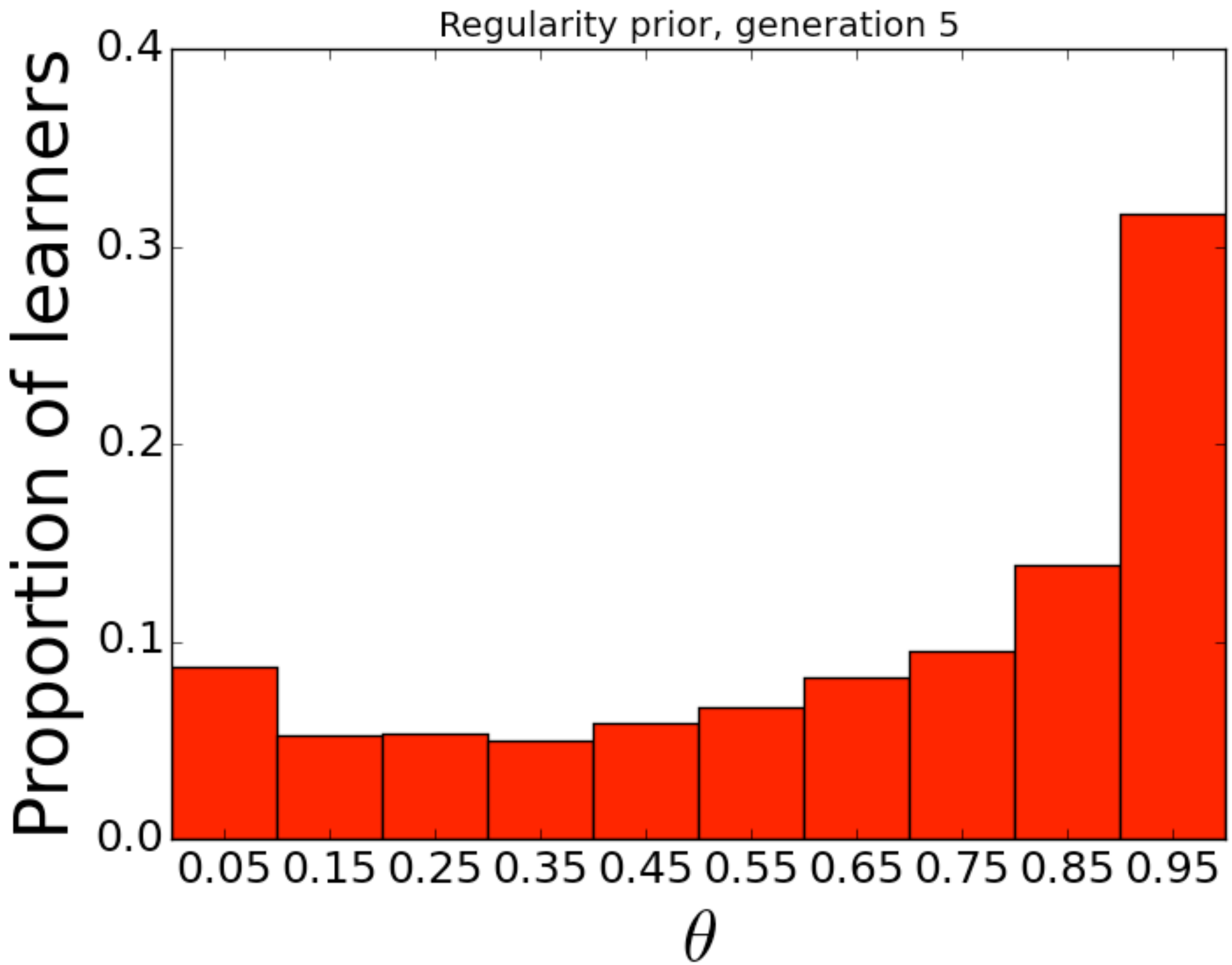
Watching the prior reveal itself



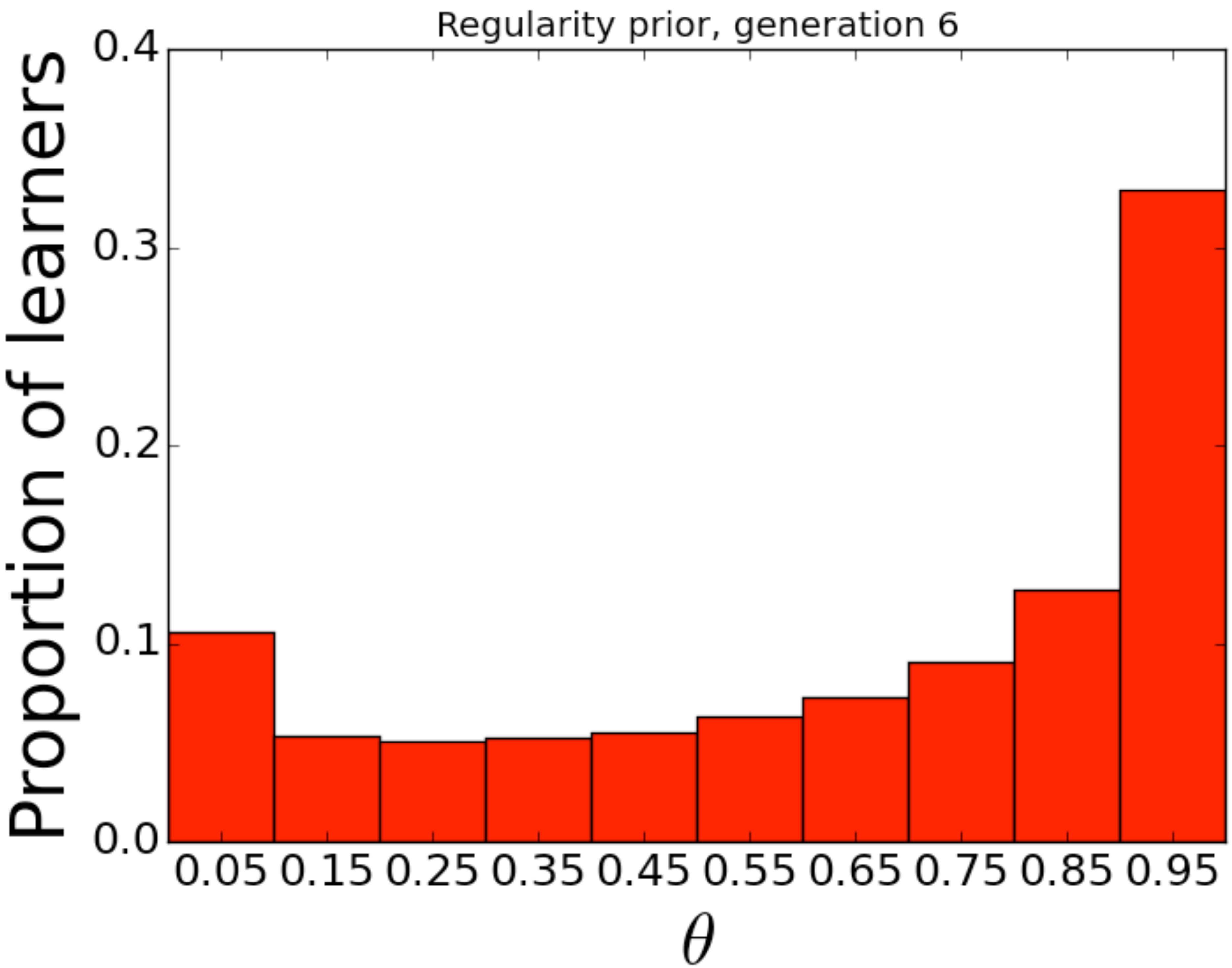
Watching the prior reveal itself



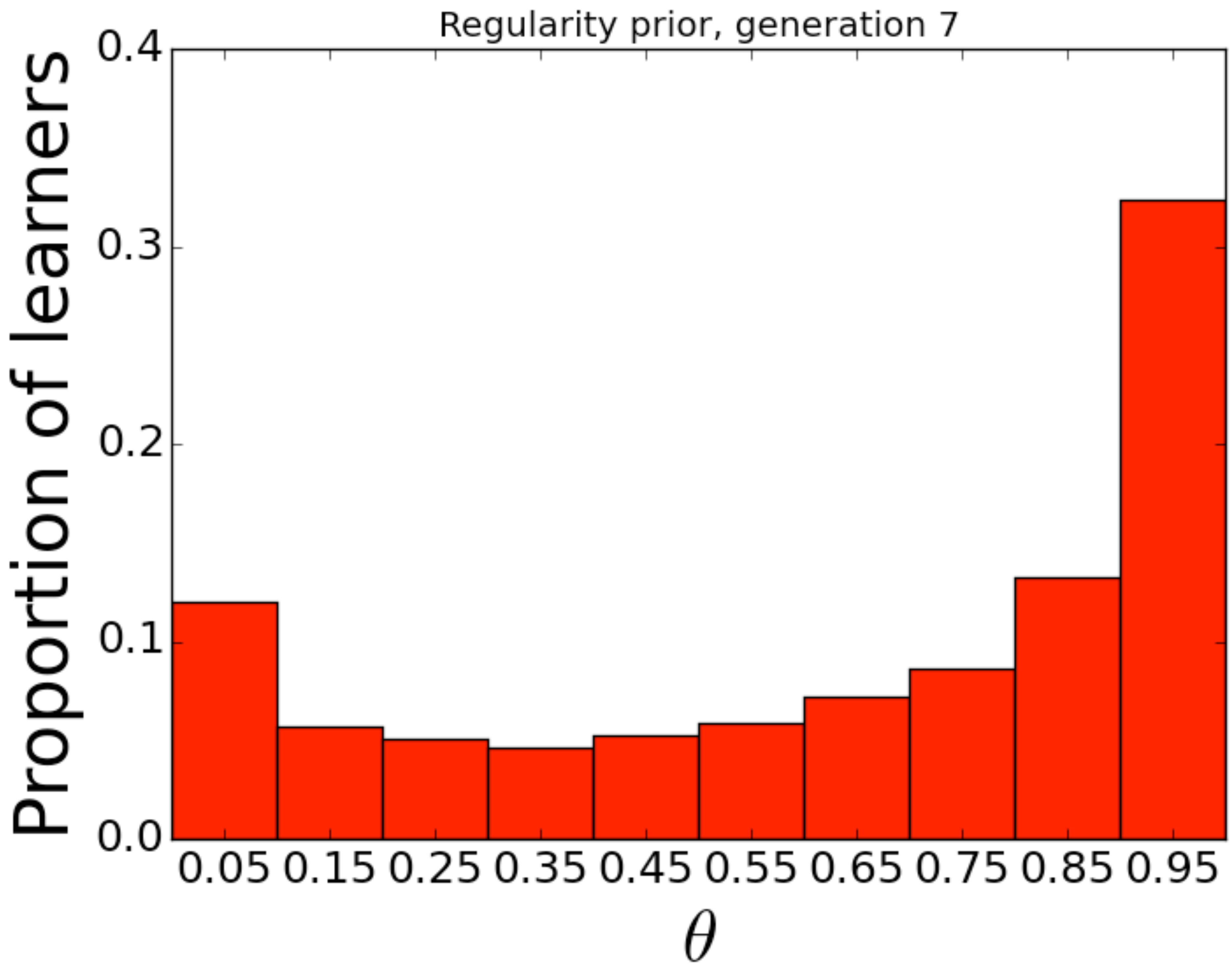
Watching the prior reveal itself



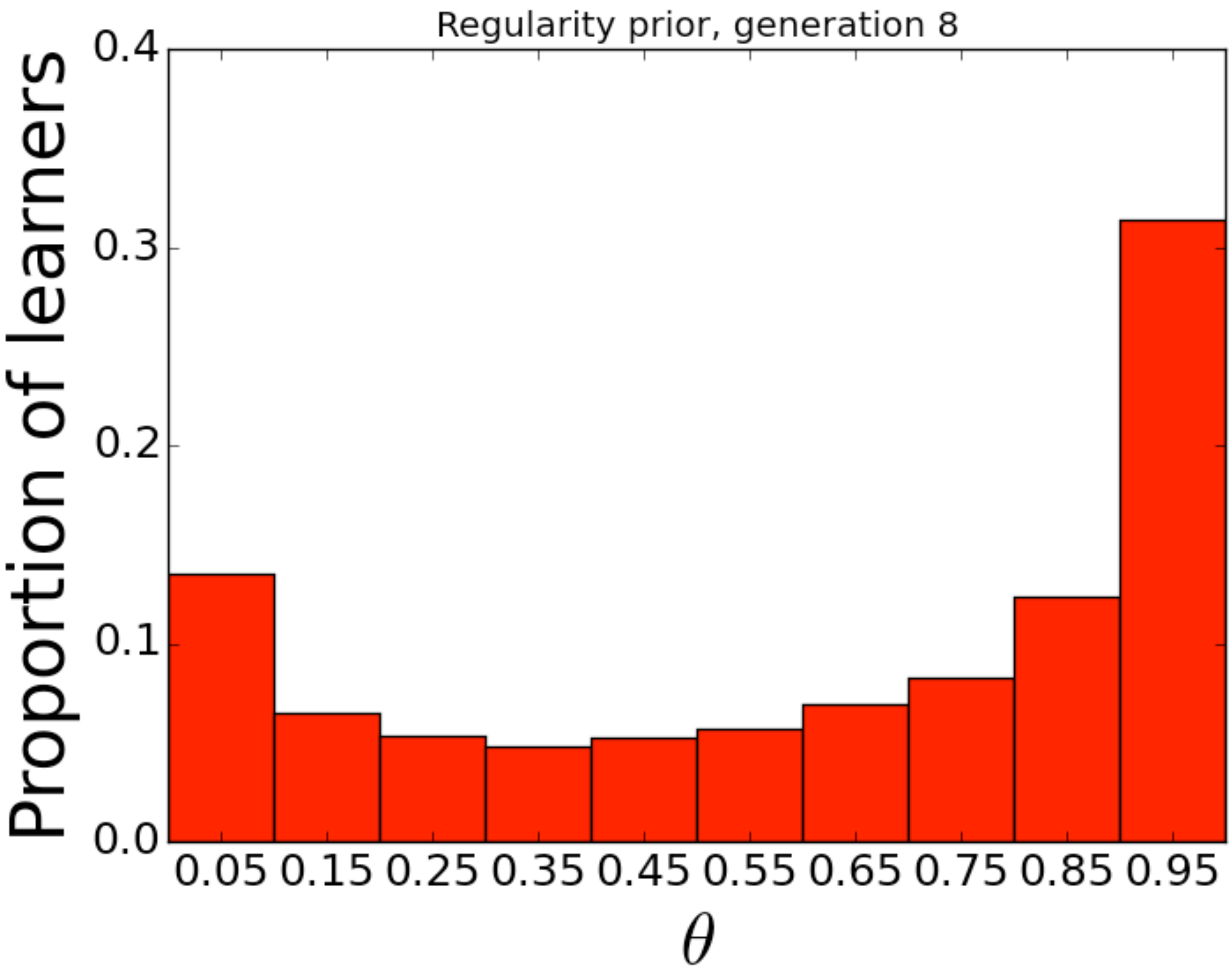
Watching the prior reveal itself



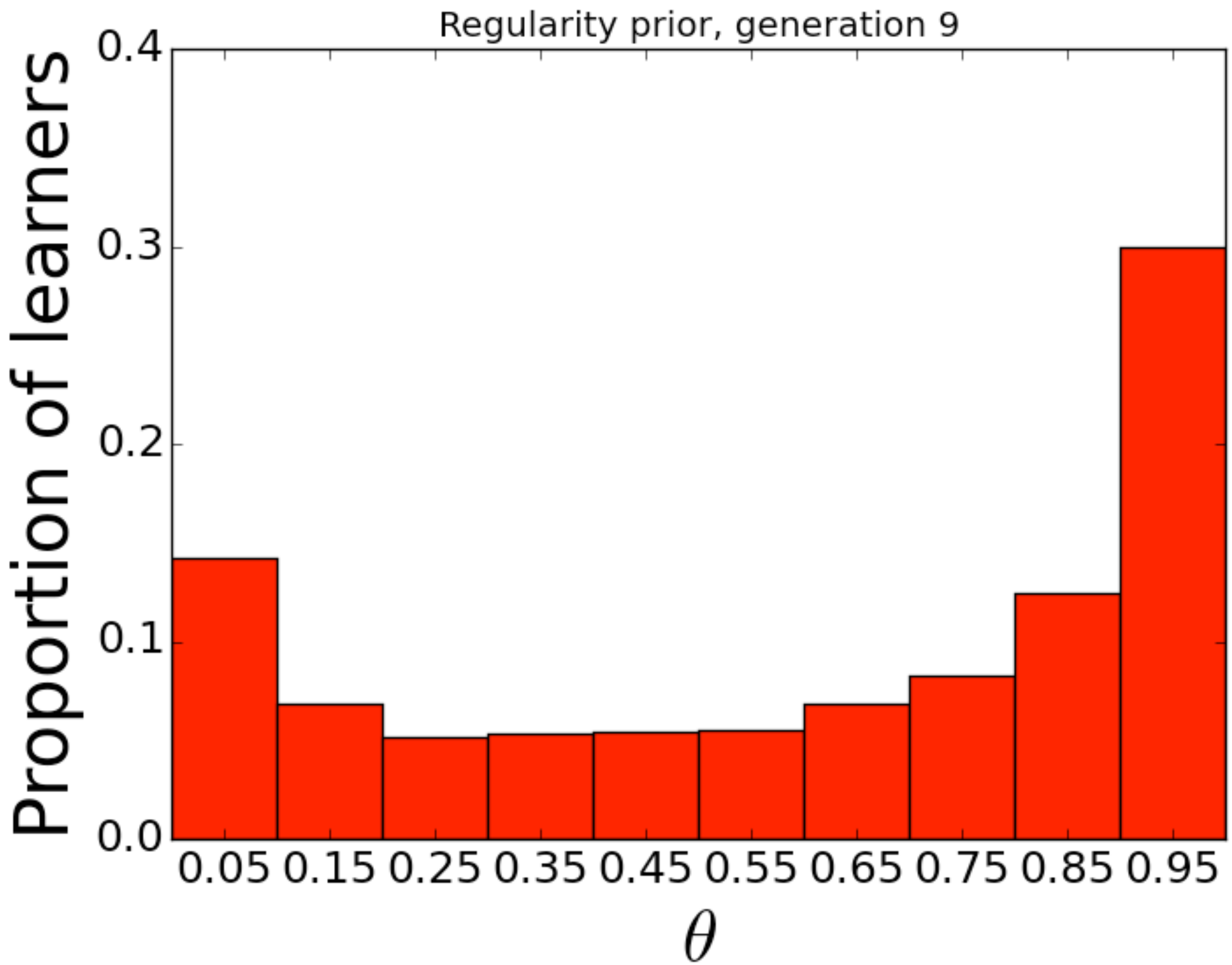
Watching the prior reveal itself



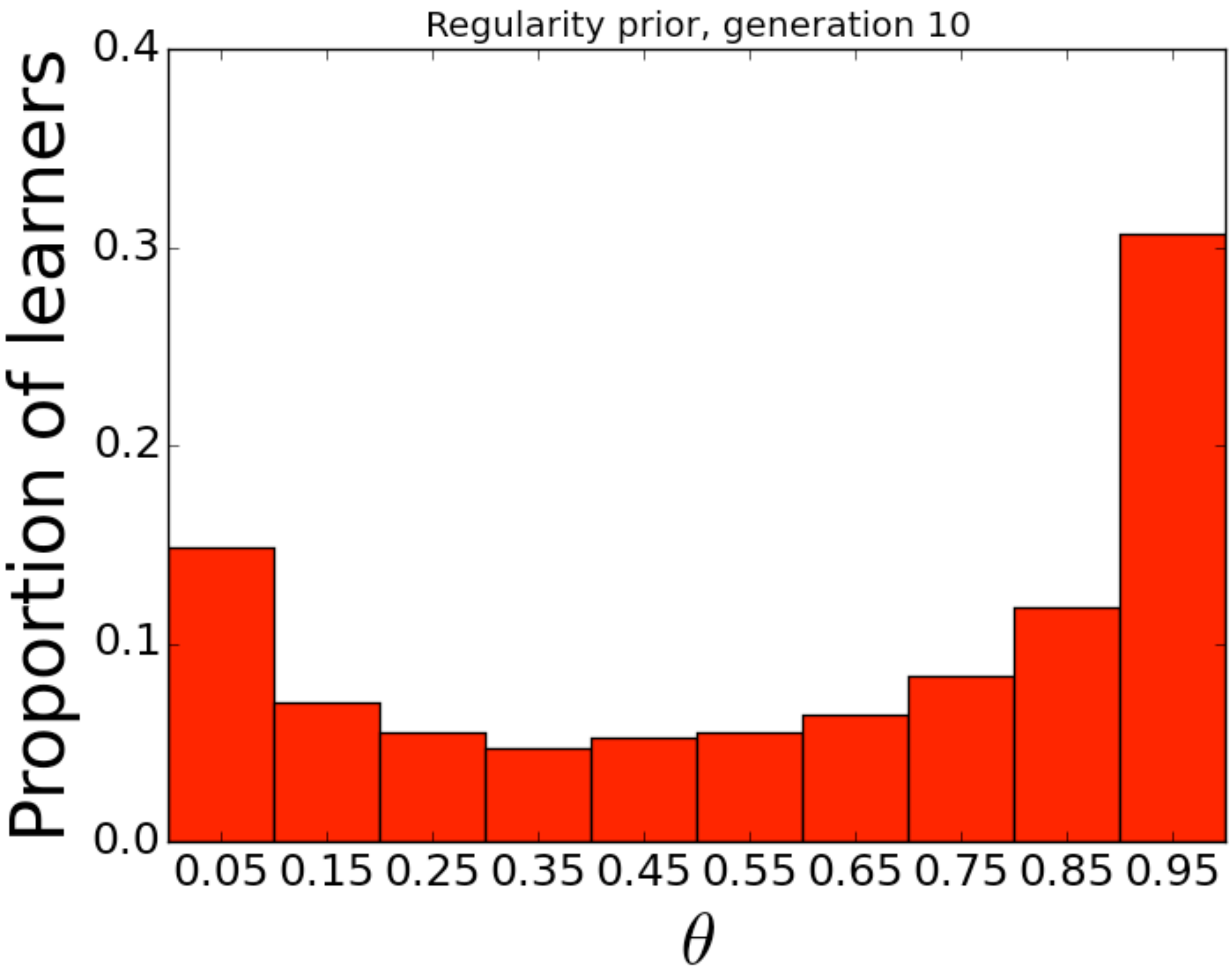
Watching the prior reveal itself



Watching the prior reveal itself



Watching the prior reveal itself



Summary and next up

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

- Bayesian learning: a nice simple way to model learning
- Make the bias of learners beautifully explicit
- **Beta-binomial model** (which is what we've been describing) allows us to model how learners respond to variability
- Two important insights:
  - If you study learning in individuals, data can obscure the prior
  - The prior can reveal itself over iterated learning
- Tuesday: lab on iterated Bayesian learning
  WARNING: get started in advance!
- Thursday: Dr Jennifer Culbertson, more beta-binomial

Reading

Hudson Kam, C., & 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.