# The Language Organism
# Lecture 10: Iterated Bayesian Learning

Simon Kirby

simon@ling.ed.ac.uk

# Previously...

# Previously...

- We uncovered the importance of the *bottleneck* on cultural transmission
- It drives the evolution of structure because only structured languages can be stably transmitted through a bottleneck (without a bottleneck, language could stay holistic)
- This is a case of adaptation for learnability by a culturally evolving language

# Previously...

- We uncovered the importance of the *bottleneck* on cultural transmission
- It drives the evolution of structure because only structured languages can be stably transmitted through a bottleneck (without a bottleneck, language could stay holistic)
- This is a case of adaptation for learnability by a culturally evolving language
- Earlier in the course, we looked at adaptation to *bias* (e.g. one-to-one constructor bias leading to optimal languages)
- Argued that this means that cultural evolution can potentially explain language structure (biological evolution by natural selection isn't the only possible explanation)

# But...

- Two possible problems:

    1. What is this thing called bias?
       How can we measure it?
       What is the bias of the syntactic learner from the last lecture, for example?

    2. If language ends up reflecting our learning biases, where does learning bias come from?

# But...

- Two possible problems:

    1. What is this thing called bias?
       How can we measure it?
       What is the bias of the syntactic learner from the last lecture, for example?

    2. If language ends up reflecting our learning biases, where does learning bias come from?

- Could we simply be restating the Chomskyan position in different terms?

  Language mirrors biologically provided innate constraints

  vs.

  Language mirrors biologically provided learning bias

# We need a more general model

- Ideally, we'd like to be able to mix up a bunch of simple ingredients and work out what language should look like after cultural evolution has run for some time:
  - BIAS (i.e. what agents are born with)
  - LANGUAGE MODEL (i.e. set of possible languages, set of possible data)
  - BOTTLENECK (i.e. how much data a learner sees)
  - POPULATION MODEL (e.g. diffusion chain, closed group etc.)
  - OTHER FEATURES OF CULTURAL TRANSMISSION (e.g. errors)

# Towards a more general model of learning bias: a medical quiz

# Towards a more general model of learning bias: a medical quiz

- Your friend coughs.  Is this cough caused by:

  a. Lung cancer

  b. A cold

  c. Athlete’s foot

# Towards a more general model of learning bias: a medical quiz

- Your friend coughs. Is this cough caused by:

  a. Lung cancer

  b. A cold

  c. Athlete's foot

- Resolving this question requires you to draw on two probabilities:
  - How likely is it that someone with the illness in question would exhibit that symptom?
  - How common is each illness?

# Likelihood of symptoms given illnesses

# Likelihood of symptoms given illnesses

**Lung cancer**: coughing is very likely, if you have lung cancer

**A cold**: coughing is very likely, if you have a cold

**Athlete's foot**: coughing is very very unikely to be caused by athlete's foot

# Likelihood of symptoms given illnesses

**Lung cancer**: coughing is very likely, if you have lung cancer

**A cold**: coughing is very likely, if you have a cold

**Athlete's foot**: coughing is very very unikely to be caused by athlete's foot

- If all we care about are the likelihood of the symptoms given each illness, we would conclude that your friend either has lung cancer or a cold

# Probability of illnesses

# Probability of illnesses

**Lung cancer**: is very rare

**A cold**: the common cold is very common

**Athlete’s foot**: is very common (let’s say)

# Probability of illnesses

**Lung cancer**: is very rare

**A cold**: the common cold is very common

**Athlete's foot**: is very common (let's say)

- If all we care about are the prevalances of each illness, we would conclude that your friend either has a cold or athlete's foot

# Probability of illnesses

**Lung cancer**: is very rare

**A cold**: the common cold is very common

**Athlete's foot**: is very common (let's say)

- If all we care about are the prevalances of each illness, we would conclude that your friend either has a cold or athlete's foot
- But you didn't conclude this: you brought these two quantities together in a smart way. How did you do it?

# The Bayesian approach

# The Bayesian approach

- What you're trying to figure out is the probability that your friend has a particular illness, given the symptoms they are exhibiting. We call this quantity:

$$P(illness|symptoms)$$

# The Bayesian approach

- What you're trying to figure out is the probability that your friend has a particular illness, given the symptoms they are exhibiting. We call this quantity:

$$P(illness|symptoms)$$

- We are trying to work this out based on two quantities which we know (roughly):

# The Bayesian approach

- What you're trying to figure out is the probability that your friend has a particular illness, given the symptoms they are exhibiting. We call this quantity:

$$P(illness|symptoms)$$

- We are trying to work this out based on two quantities which we know (roughly):
  - The likelihood of exhibiting a particular symptom given that you have a certain illness

$$P(symptoms|illness)$$

# The Bayesian approach

- What you're trying to figure out is the probability that your friend has a particular illness, given the symptoms they are exhibiting. We call this quantity:

$$P(illness|symptoms)$$

- We are trying to work this out based on two quantities which we know (roughly):
  - The likelihood of exhibiting a particular symptom given that you have a certain illness

$$P(symptoms|illness)$$

  - The prior probability of each illness

$$P(illness)$$

# Bayes' rule

# Bayes' rule

- Bayes' rule provides a convenient way of expressing the quantity we want to know in terms of the quantities we already know:

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

# Bayes' rule

- Bayes' rule provides a convenient way of expressing the quantity we want to know in terms of the quantities we already know:

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- Or, in full:

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

# Breaking it down

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

# Breaking it down

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

$P(illness|symptoms)$

- The thing we want to know is called the **posterior**

# Breaking it down

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

$P(illness|symptoms)$

- The thing we want to know is called the **posterior**

$P(symptoms|illness)$

- The probability of a particular set of symptoms given that you have a specific illness is called the **likelihood**

# Breaking it down

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

$P(illness|symptoms)$

- The thing we want to know is called the **posterior**

$P(symptoms|illness)$

- The probability of a particular set of symptoms given that you have a specific illness is called the **likelihood**

$P(illness)$

- The probability that you have a particular illness, independent of whatever symptoms you are exhibiting, is called the **prior**

# Breaking it down

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

$P(illness|symptoms)$

- The thing we want to know is called the **posterior**

$P(symptoms|illness)$

- The probability of a particular set of symptoms given that you have a specific illness is called the **likelihood**

$P(illness)$

- The probability that you have a particular illness, independent of whatever symptoms you are exhibiting, is called the **prior**

$P(symptoms)$

- The term on the bottom (the probability of the symptoms independent of illness) is actually not very interesting to us, since it is the same for all illnesses.

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness
- If a particular illness has low prior probability, we need some really convincing evidence to make us believe it to be true

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness
- If a particular illness has low prior probability, we need some really convincing evidence to make us believe it to be true
  - imagine if your friend was coughing blood and having seizures

# It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness
- If a particular illness has low prior probability, we need some really convincing evidence to make us believe it to be true
  - imagine if your friend was coughing blood and having seizures

# Errr... hello... isn't this a course about language?

# Errr... hello... isn't this a course about language?

- In the medical example, we were trying to use evidence provided by symptoms to **learn** (or infer) what underlying illness your friend had

# Errr... hello... isn't this a course about language?

- In the medical example, we were trying to use evidence provided by symptoms to **learn** (or infer) what underlying illness your friend had

- What if you aren't a medic, but a child listening to utterances?

    utterances = symptoms
    languages = illnesses
    bias in favour of particular languages = prior for each illness

# Errr... hello... isn't this a course about language?

- In the medical example, we were trying to use evidence provided by symptoms to **learn** (or infer) what underlying illness your friend had

- What if you aren't a medic, but a child listening to utterances?

  utterances = symptoms
  languages = illnesses
  bias in favour of particular languages = prior for each illness

- An ideal language learner will find a way of estimating the posterior probability of each possible language given the utterances hear

# Errr... hello... isn't this a course about language?

- In the medical example, we were trying to use evidence provided by symptoms to **learn** (or infer) what underlying illness your friend had

- What if you aren't a medic, but a child listening to utterances?

  utterances = symptoms
  languages = illnesses
  bias in favour of particular languages = prior for each illness

- An ideal language learner will find a way of estimating the posterior probability of each possible language given the utterances hear

- Children probably don't calculate sums in their head while learning, but if their learning process is sensible, we can characterise it this way

# Bayesian language learning

- Evaluate hypotheses about language given some prior bias (perhaps provided by your biology) and the data that you've heard

- You want to know the **posterior** but all you have direct access to is the **prior** and the **likelihood** (assuming you know how sentences are produced from a given model of language)

- Bayes' rule provides the solution:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

# Iterate it

# Iterate it

- So, a Bayesian model of learning is handy, because it allows us to be explicit about **bias.** We can simply plug in different values for the prior and change the preferences of learners.

# Iterate it

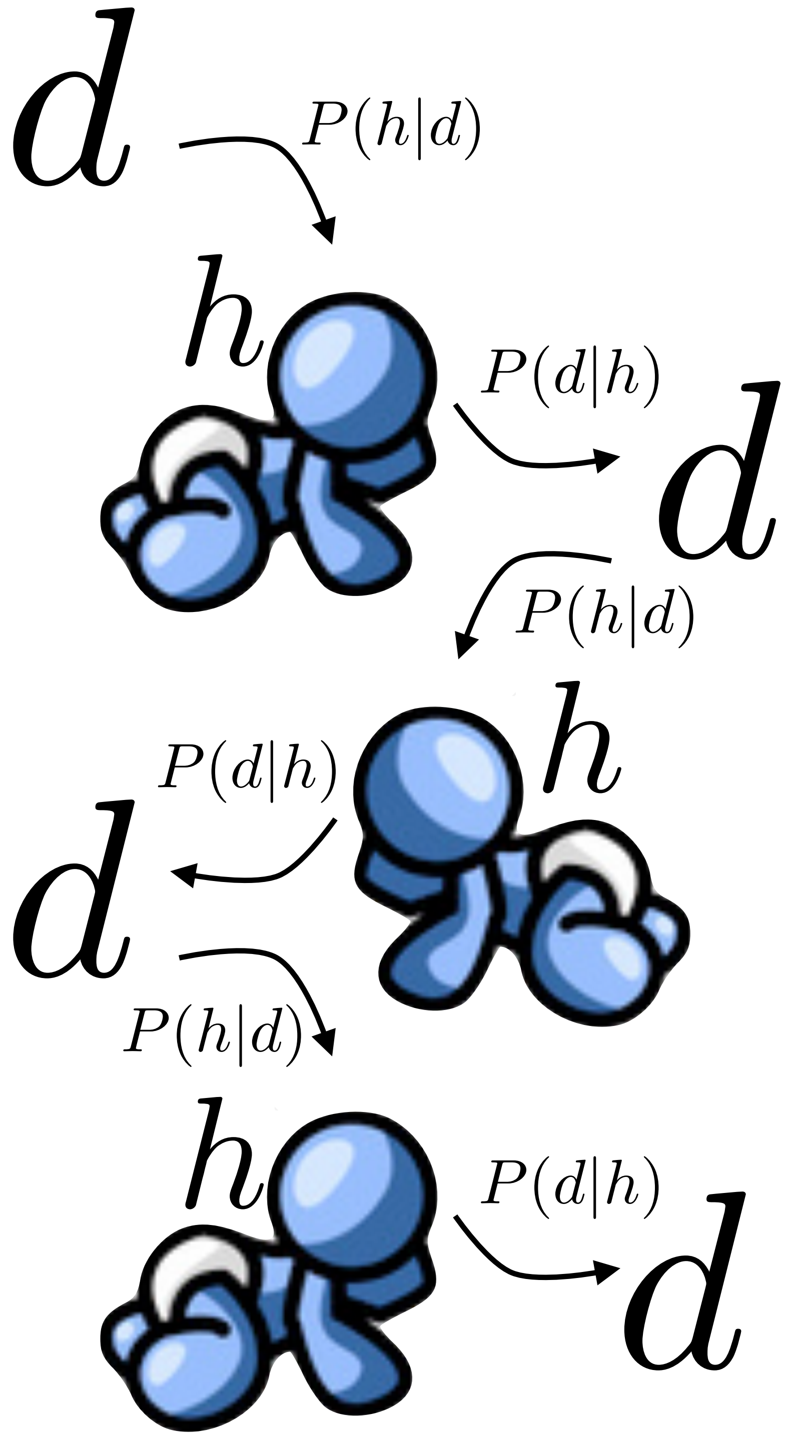
- So, a Bayesian model of learning is handy, because it allows us to be explicit about **bias.** We can simply plug in different values for the prior and change the preferences of learners.

- Now we want to know what happens in a cultural-evolutionary context.

- How does having particular bias affect the outcome of cultural evolution given particular bottlenecks, levels of noise (error) on production, and so on?

- We can put it in an iterated learning model

# Iterated Bayesian Learning




What will happen to *h* over time?

# First results (Griffiths & Kalish 2007)

# First results (Griffiths & Kalish 2007)

- Try out different models of language, different bottlenecks, different amounts of noise
- See how the process of cultural transmission takes the prior bias of the learner and gives rise to the actual resulting patterns of language

# First results (Griffiths & Kalish 2007)

- Try out different models of language, different bottlenecks, different amounts of noise
- See how the process of cultural transmission takes the prior bias of the learner and gives rise to the actual resulting patterns of language
- Any guesses as to what they showed?

# First results (Griffiths & Kalish 2007)

- Try out different models of language, different bottlenecks, different amounts of noise

- See how the process of cultural transmission takes the prior bias of the learner and gives rise to the actual resulting patterns of language

- Any guesses as to what they showed?

  Bottleneck does nothing
  Noise does nothing
  Details of language model do nothing

- Given enough time, the end result of cultural evolution always reflects the prior bias and nothing else

# Hang on a minute...

# Hang on a minute...

- This runs completely counter to the results from our previous simulation

# Hang on a minute...

- This runs completely counter to the results from our previous simulation
- We argued that it was the bottleneck that was driving adaptation of the language
- We also argued that cultural evolution has something important to add

# Hang on a minute...

- This runs completely counter to the results from our previous simulation
- We argued that it was the bottleneck that was driving adaptation of the language
- We also argued that cultural evolution has something important to add
- If prior bias is essentially what is innate to the learner, then Griffiths & Kalish seem to be saying that the universal properties of language are just a straightforward reflection of innateness

# Hang on a minute...

- This runs completely counter to the results from our previous simulation
- We argued that it was the bottleneck that was driving adaptation of the language
- We also argued that cultural evolution has something important to add
- If prior bias is essentially what is innate to the learner, then Griffiths & Kalish seem to be saying that the universal properties of language are just a straightforward reflection of innateness
- Hmmm...

# Some subtleties in the model

# Some subtleties in the model

- Kirby, Dowman & Griffiths (2007): tried to square the Bayesian model with what we **thought** we knew about cultural evolution of language

# Some subtleties in the model

- Kirby, Dowman & Griffiths (2007): tried to square the Bayesian model with what we **thought** we knew about cultural evolution of language

- Whole thing revolves around a very subtle point

  - How do you decide, given the posterior, whether your friend has cancer, a cold or athlete's foot?

  - How do you decide, given the posterior, which language to select?

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

# Sampling vs. MAP

# Sampling vs. MAP

- There are (at least) two sensible choices:

# Sampling vs. MAP

- There are (at least) two sensible choices:
  - Sampling: given a particular distribution of probabilities, pick your hypothesis from the distribution proportionally.

    (If it's ten times more likely to be a cold than cancer, once in a while you'll tell your friend you think it's cancer)

# Sampling vs. MAP

- There are (at least) two sensible choices:
  - Sampling: given a particular distribution of probabilities, pick your hypothesis from the distribution proportionally.

    (If it's ten times more likely to be a cold than cancer, once in a while you'll tell your friend you think it's cancer)

  - MAP: given a particular distribution of probabilities, pick the best. This is called the maximum a-posteriori (MAP) hypothesis

    (If it's more likely to be a cold than cancer, tell them you think it's a cold)

# Sampling vs. MAP

- There are (at least) two sensible choices:
  - Sampling: given a particular distribution of probabilities, pick your hypothesis from the distribution proportionally.

    (If it's ten times more likely to be a cold than cancer, once in a while you'll tell your friend you think it's cancer)
  - MAP: given a particular distribution of probabilities, pick the best. This is called the maximum a-posteriori (MAP) hypothesis

    (If it's more likely to be a cold than cancer, tell them you think it's a cold)
- Griffith & Kalish (2007) were using *sampling.* Kirby et al. (2007) tried MAP.

# A simple example: regularity

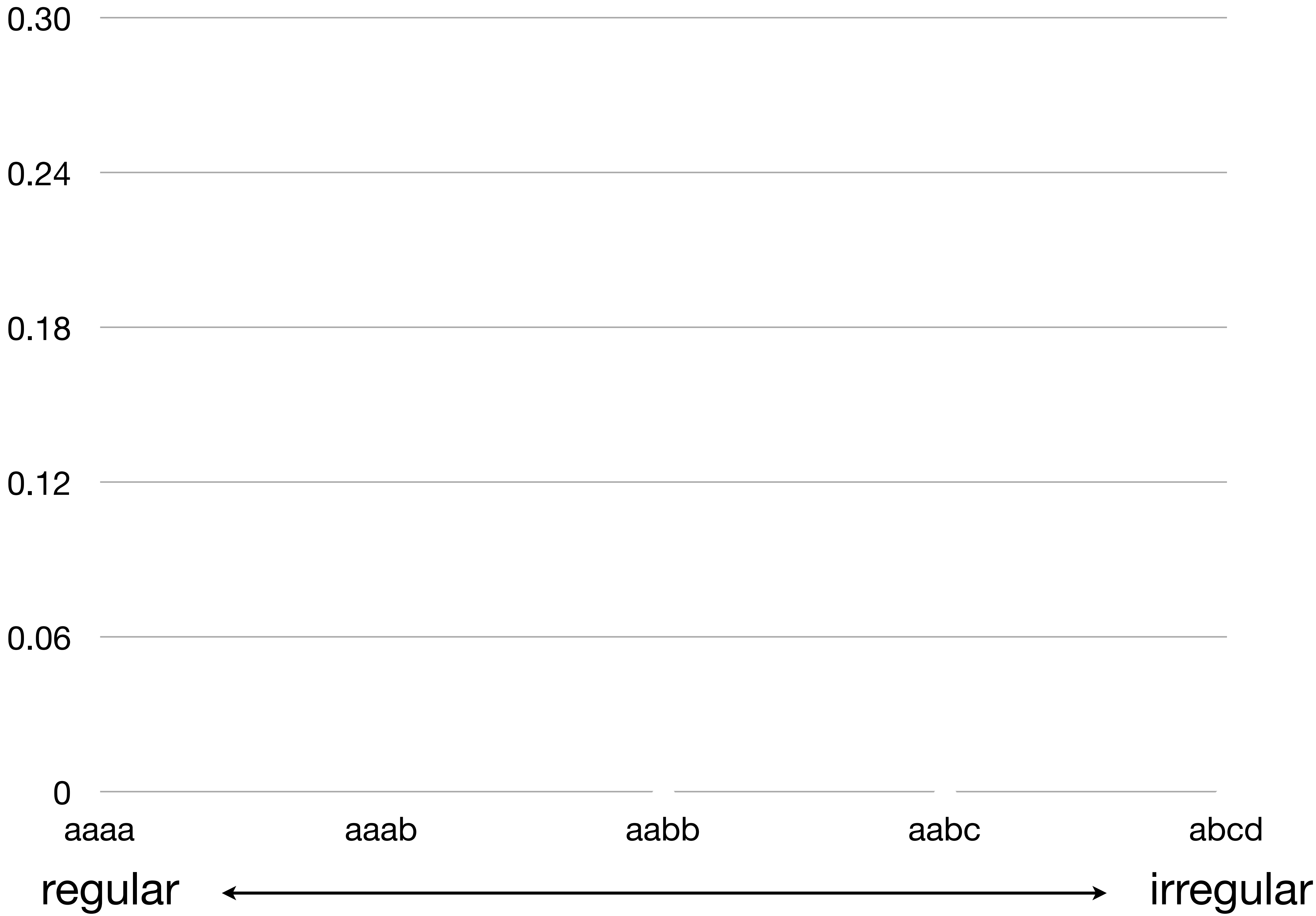
# A simple example: regularity

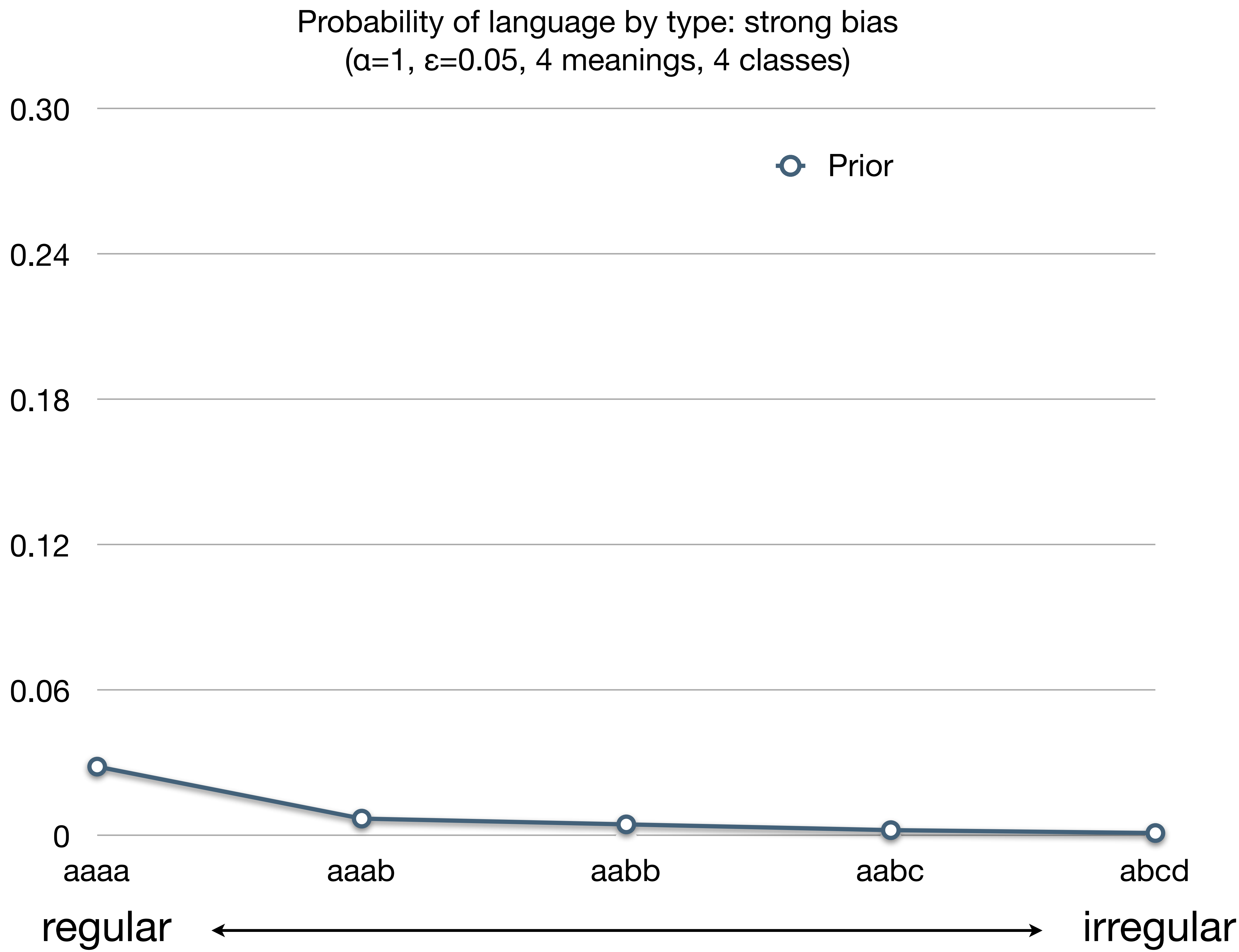
- Model language as a set of meanings
- These meanings can be expressed regularly, or irregularly

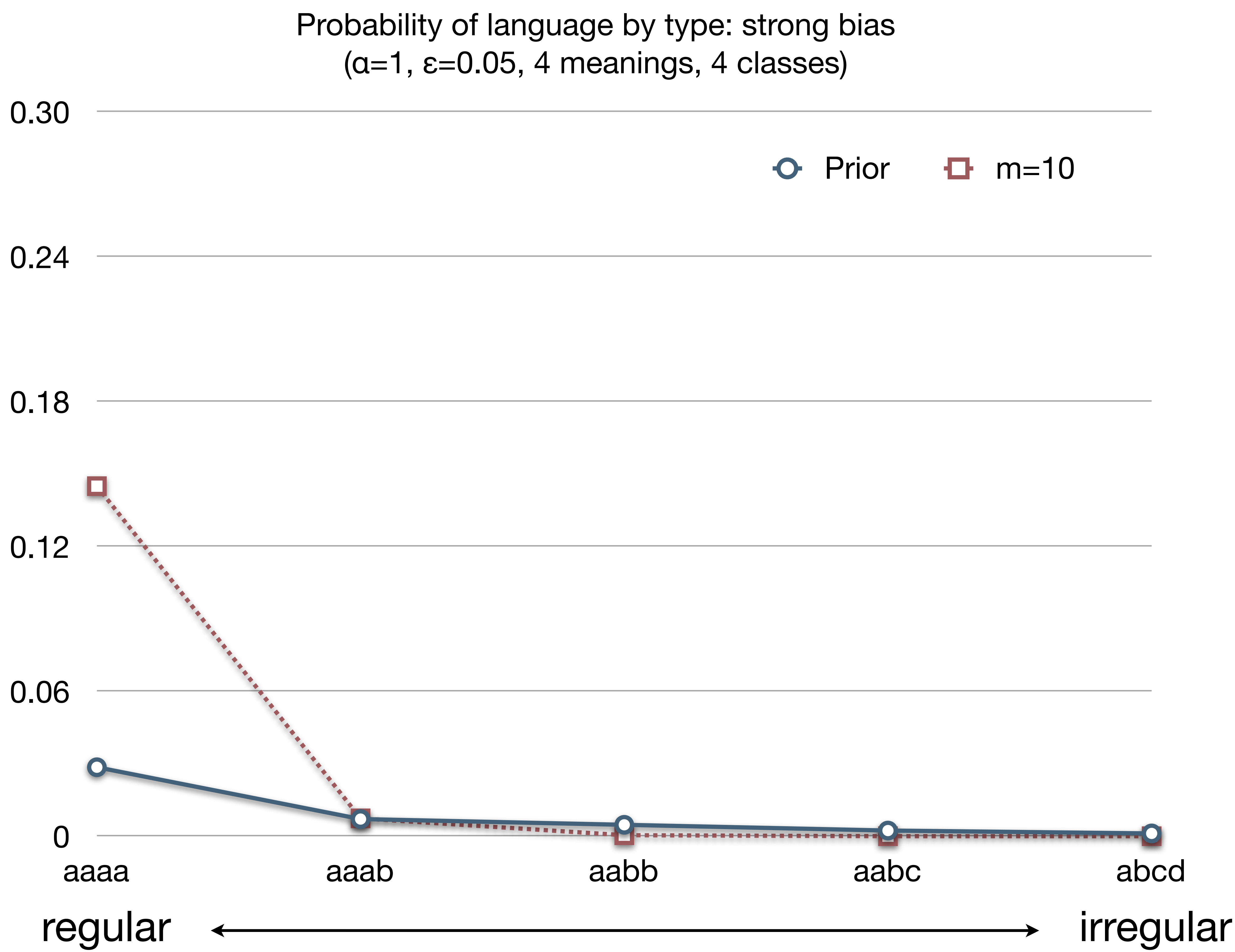
# A simple example: regularity

- Model language as a set of meanings
- These meanings can be expressed regularly, or irregularly
- Start with the assumption that there is a slight innate bias in favour of regularity
    - We can vary the strength of this bias
    - It is reasonable to assume a simple bias like this is not language-specific
- Assume learners pick the best (i.e. MAP) hypothesis. What happens?

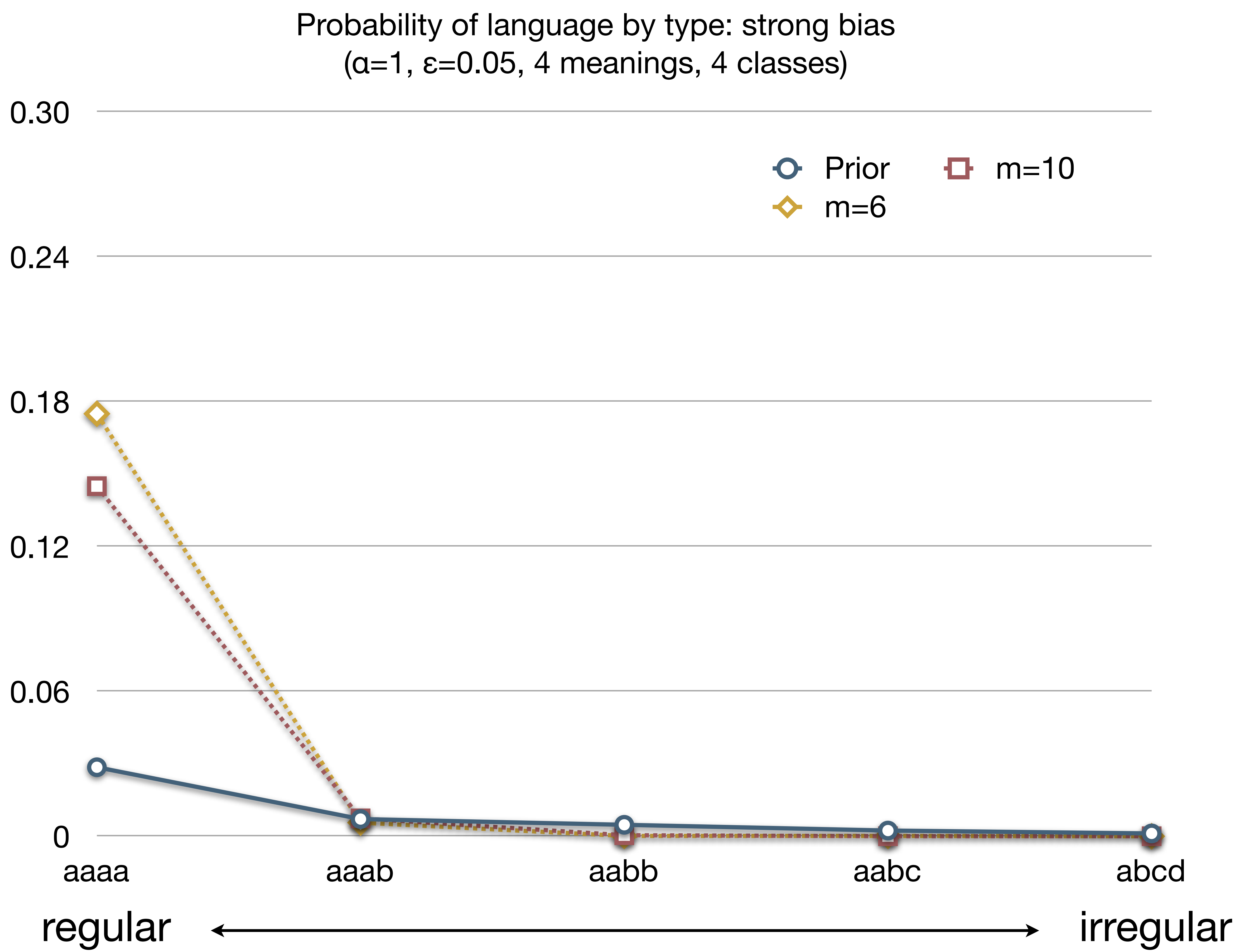
Probability of language by type: strong bias
(α=1, ε=0.05, 4 meanings, 4 classes)

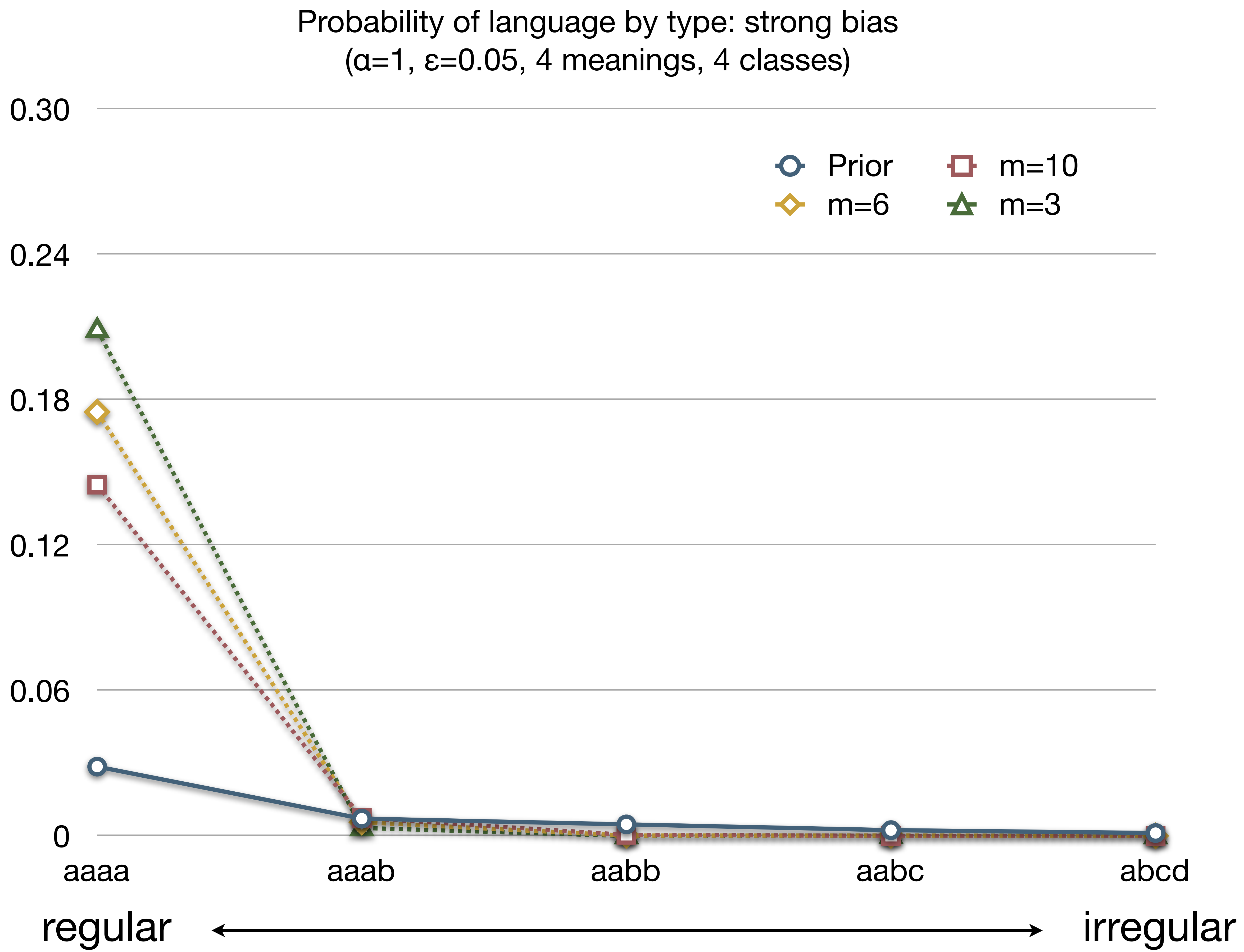



Probability of language by type: strong bias
(α=1, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
aaaa
aaab
aabb
aabc
abcd
regular
irregular

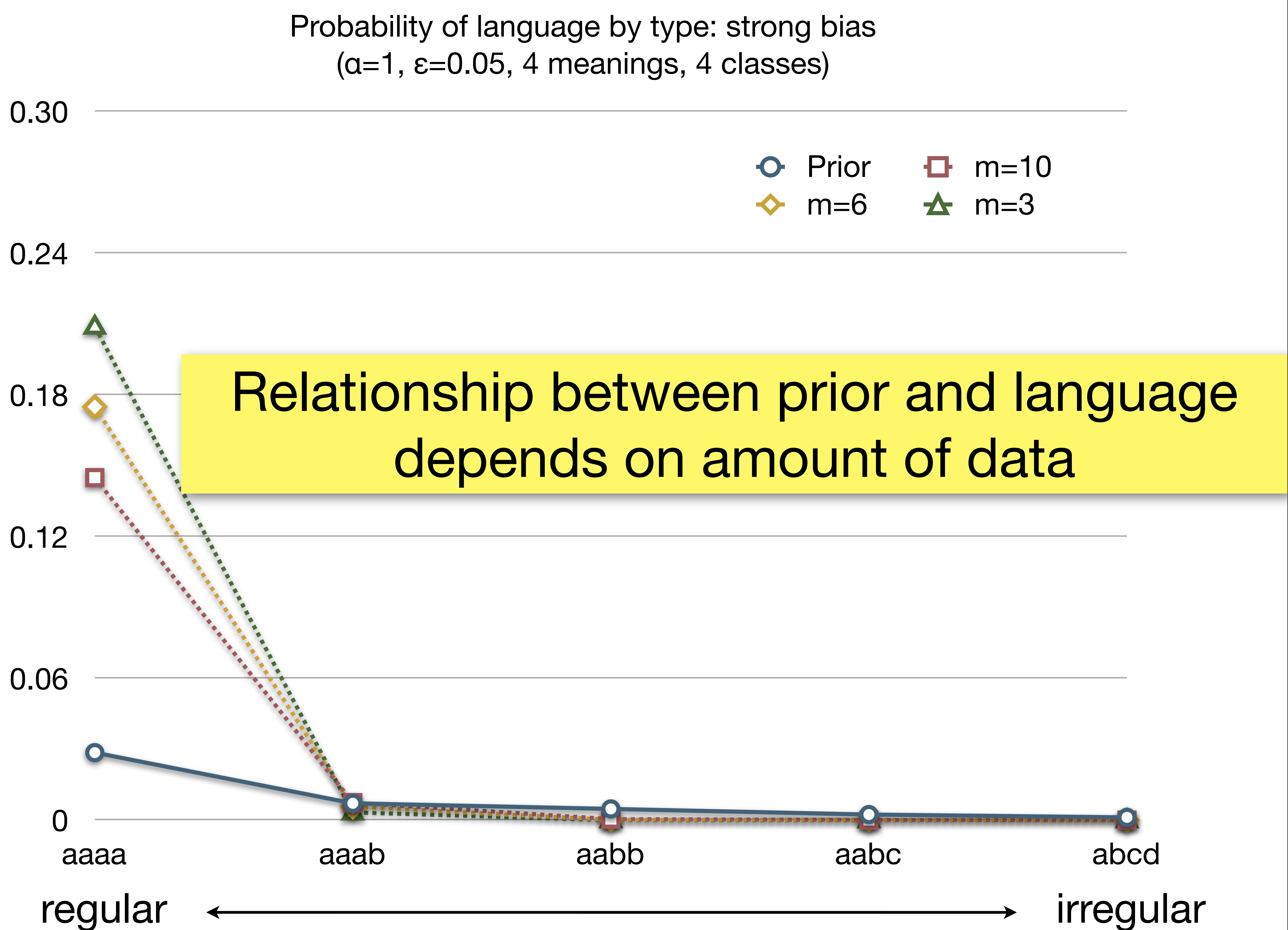


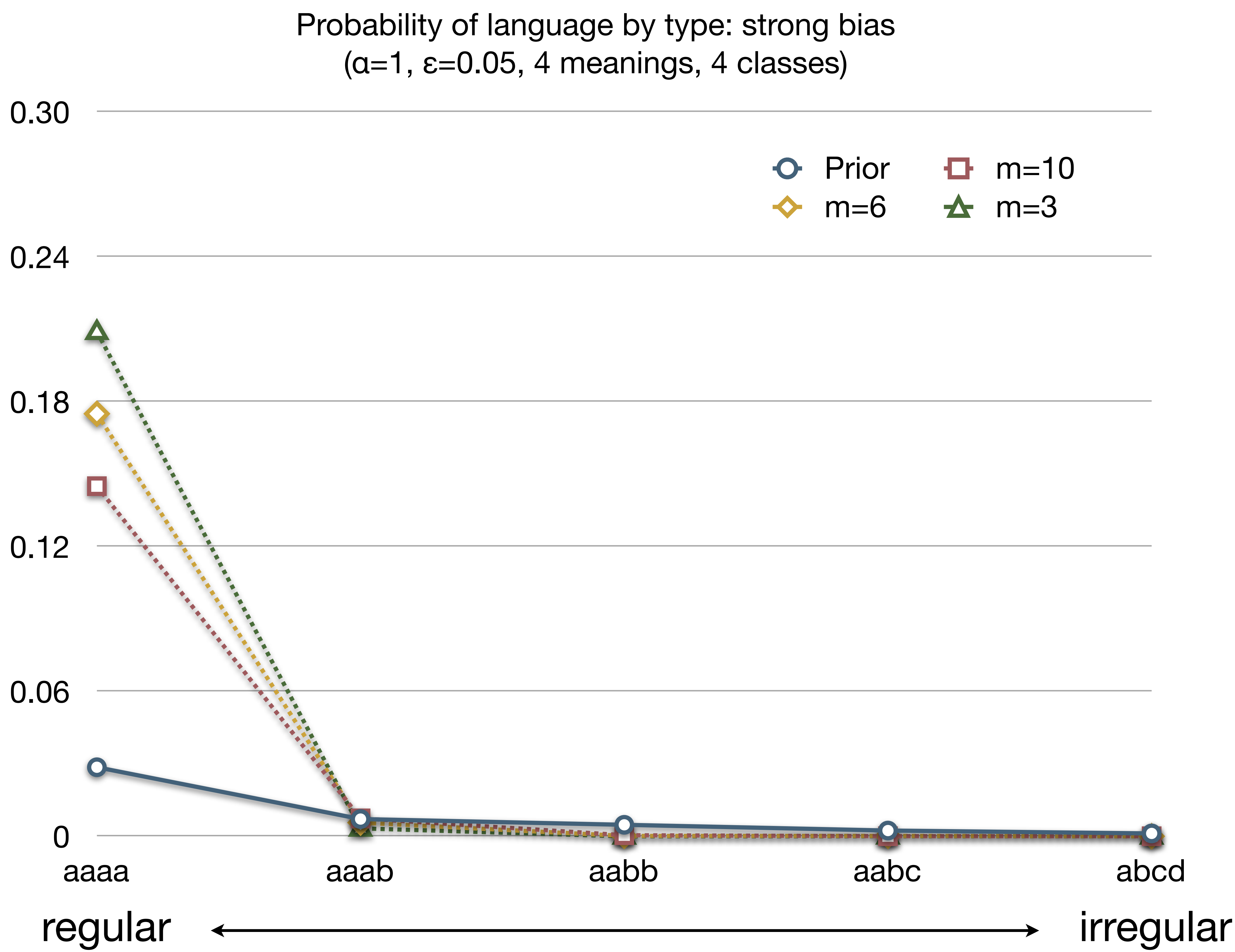
Probability of language by type: strong bias
(α=1, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=10
aaaa
aaab
aabb
aabc
abcd
regular
irregular

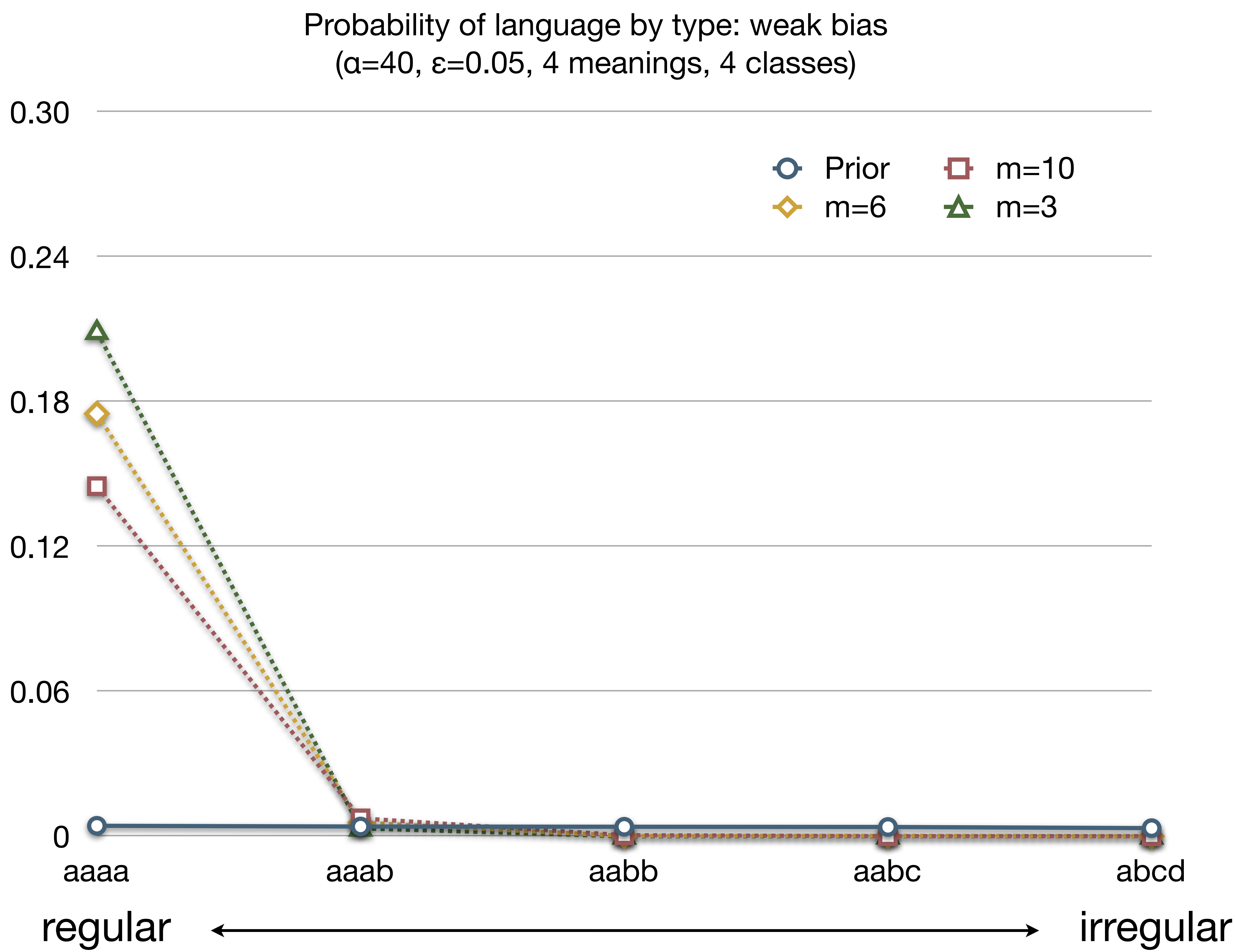

Probability of language by type: strong bias
(α=1, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=10
m=6
aaaa
aaab
aabb
aabc
abcd
regular
irregular



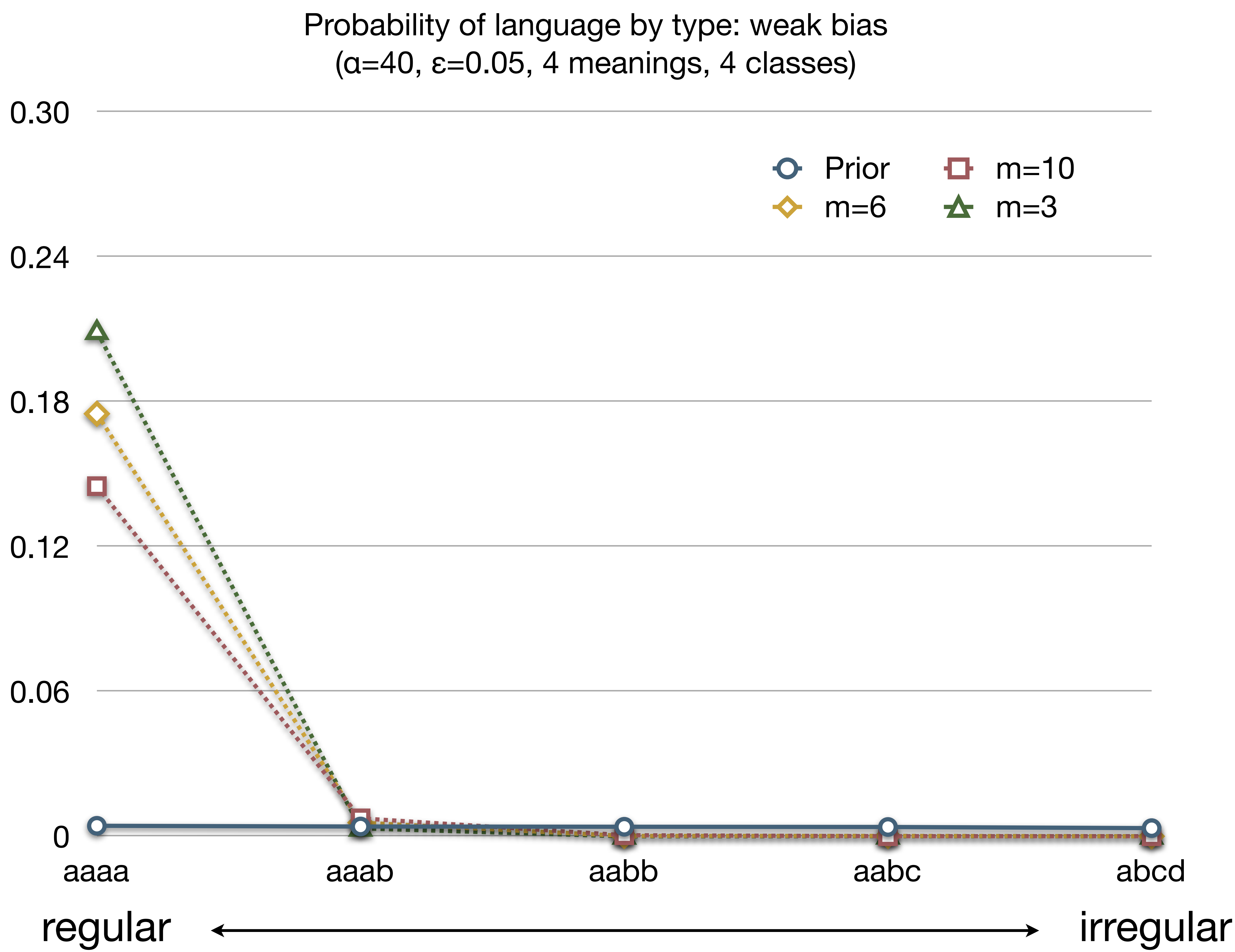
Probability of language by type: strong bias
(α=1, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=6
m=10
m=3
aaaa
aaab
aabb
aabc
abcd
regular
irregular

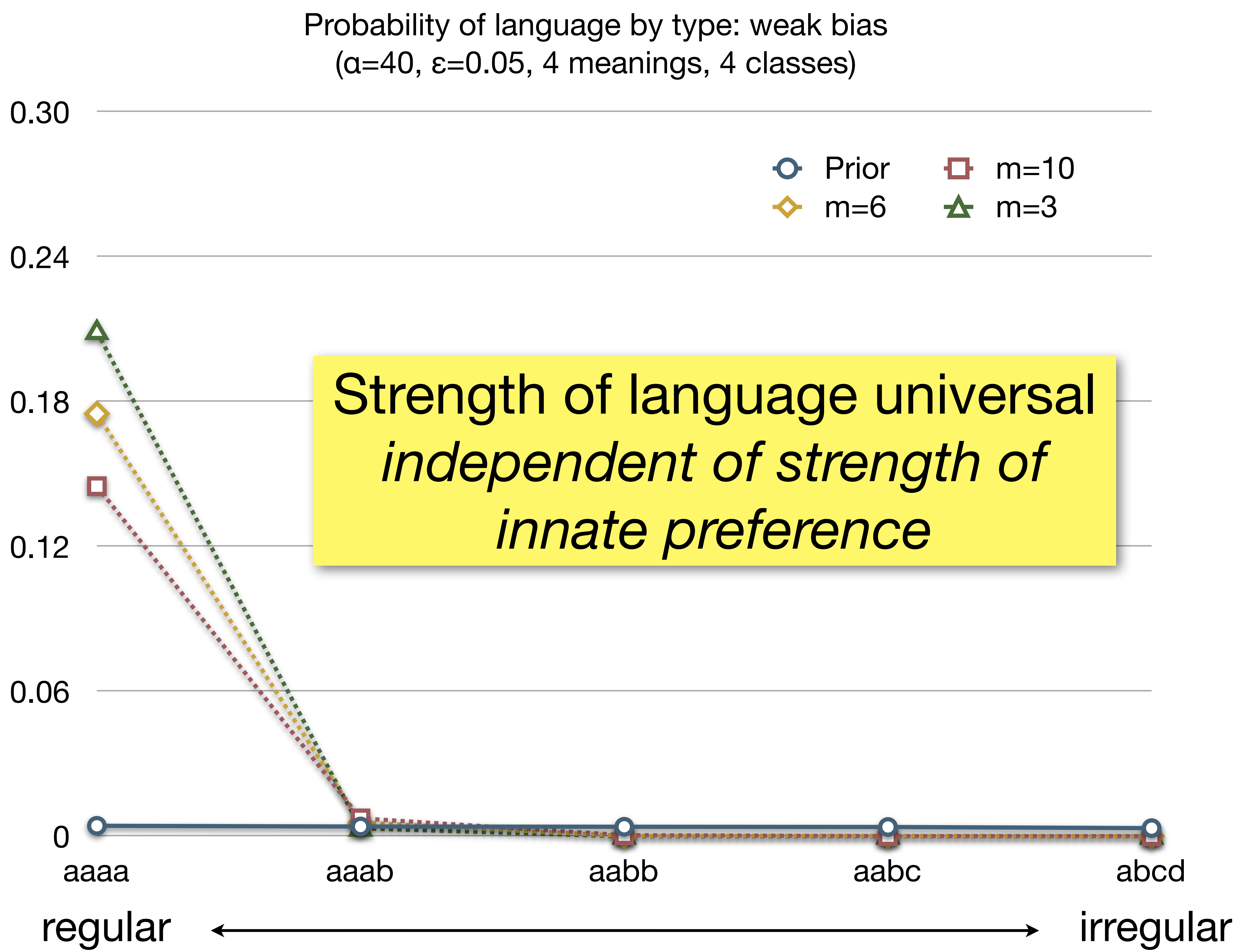

Probability of language by type: strong bias
(α=1, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=10
m=6
m=3
aaaa
aaab
aabb
aabc
abcd
regular
irregular
Relationship between prior and language depends on amount of data


Probability of language by type: strong bias
(α=1, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=10
m=6
m=3
aaaa
aaab
aabb
aabc
abcd
regular
irregular



Probability of language by type: weak bias
(α=40, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=6
m=10
m=3
aaaa
aaab
aabb
aabc
abcd
regular
irregular


Probability of language by type: weak bias
(α=40, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=10
m=6
m=3
aaaa
aaab
aabb
aabc
abcd
regular
irregular


Probability of language by type: weak bias
(α=40, ε=0.05, 4 meanings, 4 classes)
0.30
0.24
0.18
0.12
0.06
0
Prior
m=10
m=6
m=3
Strength of language universal
*independent of strength of*
*innate preference*
aaaa
aaab
aabb
aabc
abcd
regular
irregular

# Conclusions

# Conclusions

- Iterated Bayesian Learning allows us to more precisely understand the relationship between learning bias and eventual language structure

# Conclusions

- Iterated Bayesian Learning allows us to more precisely understand the relationship between learning bias and eventual language structure

- If you assume social learning is about maximising the chance of converging on what other people are doing (i.e. selecting the MAP hypothesis), then cultural evolution does a lot of work for you

# Conclusions

- Iterated Bayesian Learning allows us to more precisely understand the relationship between learning bias and eventual language structure

- If you assume social learning is about maximising the chance of converging on what other people are doing (i.e. selecting the MAP hypothesis), then cultural evolution does a lot of work for you

- Very weak innate biases are all that's needed to explain strong linguistic universals

# Conclusions

- Iterated Bayesian Learning allows us to more precisely understand the relationship between learning bias and eventual language structure
- If you assume social learning is about maximising the chance of converging on what other people are doing (i.e. selecting the MAP hypothesis), then cultural evolution does a lot of work for you
- Very weak innate biases are all that's needed to explain strong linguistic universals
- If we see universals in language, then we should not be assuming that these are hard-coded as strong-constraints in the genes