Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 3

Last class

Maximum likelihood estimation:

$$L(\theta, w_1, \ldots, w_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(w_i \mid \theta)$$

$$\theta^*_{MLE} = \arg\max_{\theta} L(\theta, w_1 \ldots w_n)$$

A guessing game

I choose a random number between 1 and 20. You need to guess it, and each time you make a guess I tell you whether your guess is higher or lower than my number. What is your strategy to guess the number as quickly as possible?

I choose a random number x between 1 and 20 **from a distribution** $p(x)$. You know p and need to guess the number. What is your strategy?

$-\log_2 p(x)$

What does log-probability mean?

Let p be a probability distribution over Ω. What is $-\log_2 p(x)$?

bits to encode

$$-\sum (\log_2 p(x))\, p(x) = H(p)$$ entropy

If you think the distribution is q: $-\log_2 q(x)$ for every x

Cross-entropy $$-\sum p(x) \log_2 q(x)$$

Another view of maximum likelihood estimation

What is the "empirical distribution?"

$w_1, \ldots, w_n$

$$\tilde{p}(x) = \begin{cases} \frac{\text{count}(w)}{n} & w \in D \\ 0 & w \notin D \end{cases}$$

Rewriting the objective function $L(\theta, w_1, \ldots, w_n)$

$$L(\theta, w_1 \ldots w_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(w_i \mid \theta) =$$

$$= \sum_{w \in \Omega} \underline{\underline{\tilde{p}(x)}} \log p(w_i \mid \theta) = -CE(\tilde{p}, p_\theta)$$

$$\min CE(\tilde{p}, p_\theta)$$

Cross-entropy

What is the definition of cross-entropy?

$$- \sum_{x \in \Omega} \left(\log_2 q(x) \right) p(x) = CE(p, q)$$

Likelihood maximisation

By doing maximum likelihood maximisation we:

- Choose the parameters that make the data most probable,

  or, from an information-theoretic perspective:

- Choose the parameters that make the encoding of the data most succinct (bit-wise),

  in other words, we

- Minimize the cross-entropy between the empirical distribution and the model we choose.

Types of Models

It is often the case that we discuss a model $p(x \mid \theta)$

Really, in NLP, you are interested in predicting some $y(x)$

Therefore, you need $p(x, y \mid \theta)$. Estimation is the same when both x and y are in the dataset. Later we will learn about incomplete data

In some cases you model also $p(y \mid x, \theta)$ (e.g. neural networks, log-linear models).

This gives the generative vs. discriminative model distinction

Types of Objectives

We showed an example of deriving the log-likelihood solution for a simple model

One can have more complex objective functions, and the principle would be the same

You just might not have a closed-form solution (e.g. with deep learning, log-linear models, etc.)

You need to apply an *optimisation* algorithm – more on that later

Some history

- History: 1700s. Seminal ideas due to Thomas Bayes and Pierre-Simon Laplace

Use of Bayesian Learning in NLP

Very often used in the context of *unsupervised* learning. Why?

- Hard to beat discriminative methods in the supervised case (log-linear models, deep learning, etc.)
- Flexible framework for latent variables
- Priors play much more important role in the unsupervised setting

See more discussion in the reading material

Bayes' rule

What is Bayes' rule? If you have X and Y

$$p(X=x \mid Y=y) = \frac{p(Y=y \mid X=x)\, p(X=x)}{p(Y=y)}$$

by chain rule

Reminder: What does Statistics do? Invert the relationship between model and data.

Bayes' rule does the same with random variables.

Define our "X" to be θ

Define our "Y" to be the data

Bayes' rule

What is Bayes' rule?

Reminder: What does Statistics do? Invert the relationship between model and data.

Bayes' rule does the same with random variables.

What if our model parameters were one random variable and our data were another random variable?

Prior beliefs about models

We have a parameter space Θ and prior beliefs $p(\theta)$.

Our θ is now a random variable.

From the chain rule: $p(w, \theta) = p(\theta)p(w|\theta)$

"prior"

$$p(\theta|w) = \frac{p(w|\theta)\,p(\theta)}{p(w)}$$

the model

the "likelihood"

Posterior inference

likelihood ← prior

$$p(\theta \mid w) = \frac{p(w \mid \theta)p(\theta)}{p(w)}$$

basic posterior inference

$$\overline{\overline{p(\theta|w)}}$$

$$p(w) = \int_\theta p(w|\theta)\, p(\theta)\, d\theta$$

$$\int_\theta p(\theta|w)\, d\theta = 1$$

$$\int_\theta p(w|\theta)\, p(\theta) / \underline{\underline{p(w)}}\, d\theta = 1$$

$$p(w) = \int_\theta p(w|\theta)\, p(\theta)\, d\theta$$

Priors

Our prior beliefs are considered in inference. There is no “correct” prior.

Is that a good or bad thing?

Priors

Our prior beliefs are considered in inference. There is no “correct” prior.

Is that a good or bad thing?

- Frequentists: probability is the frequency of an event
- Bayesians: probability denotes the state of our knowledge about an event
    - Subjectivists: probability is a personal belief
    - Objectivists: minimise human’s influence on decision making
- In practice: NLP use of Bayesian theory is largely driven by computation

Back to pre-historic languages

Language with two words: “argh” and “blah”

Our Ω is $\{\text{argh}, \text{blah}\}$.

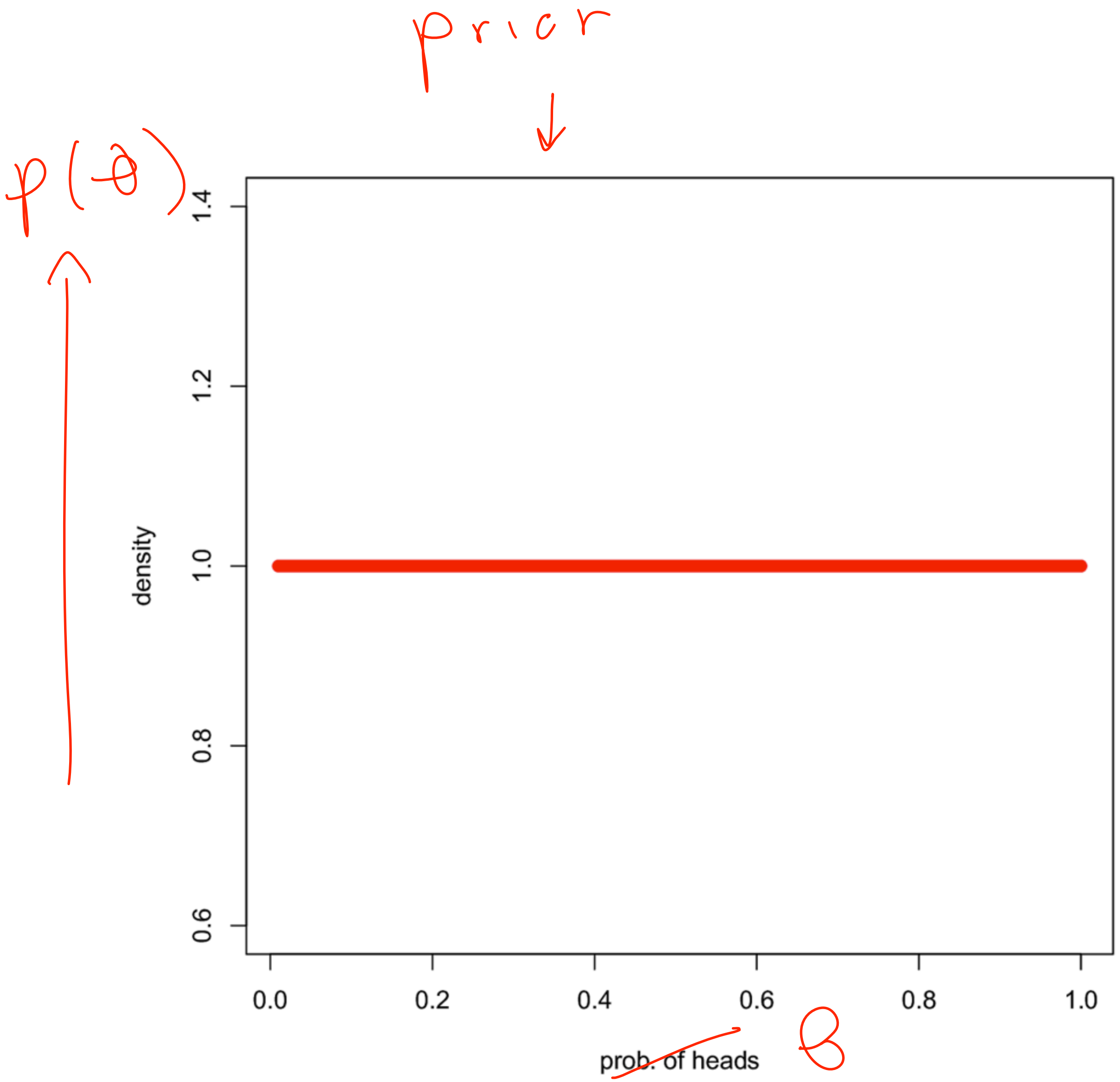
Our Θ is $[0, 1]$.

Define $I(w) = 1$ if $w = \text{argh}$ and 0 if $w = \text{blah}$.

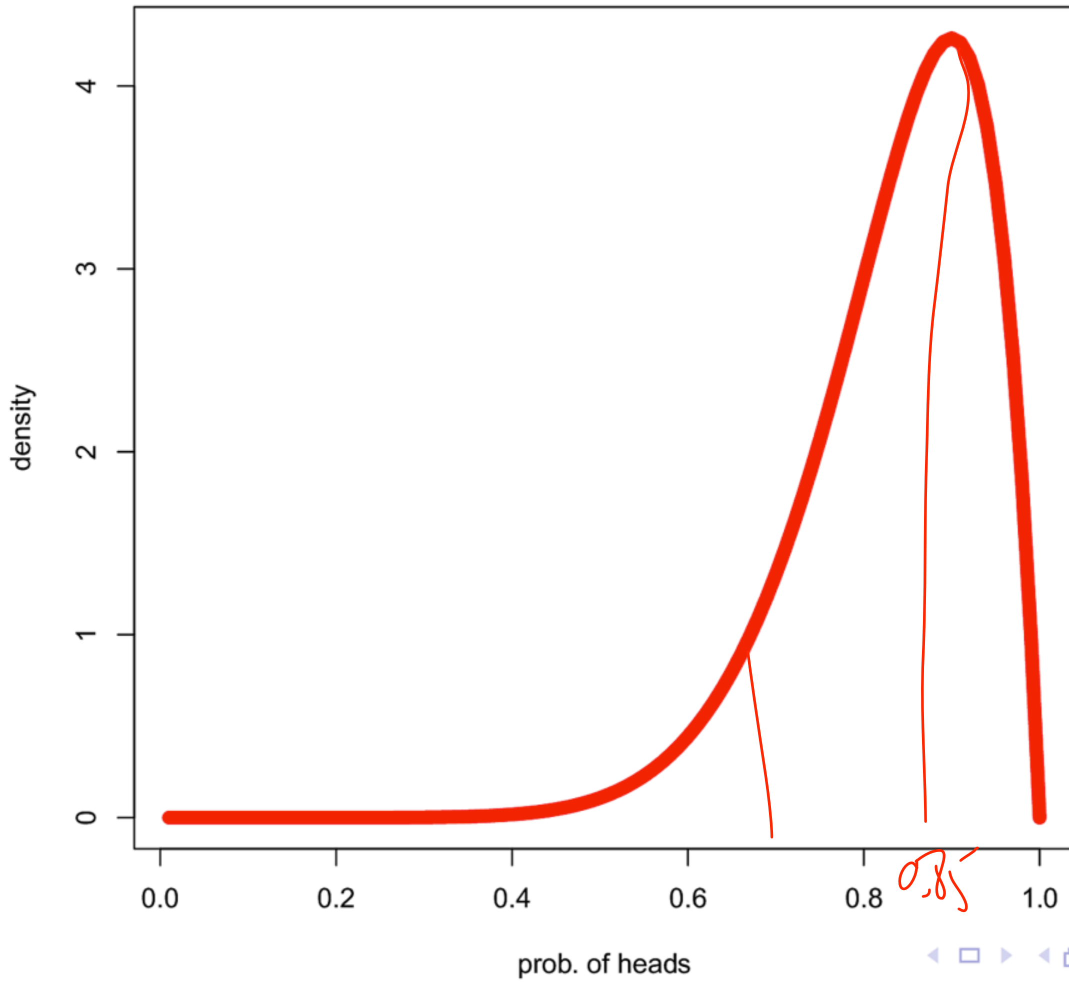
Then, $p(w|\theta) = \theta^{I(w)}(1-\theta)^{(1-I(w))}$.

likelihood

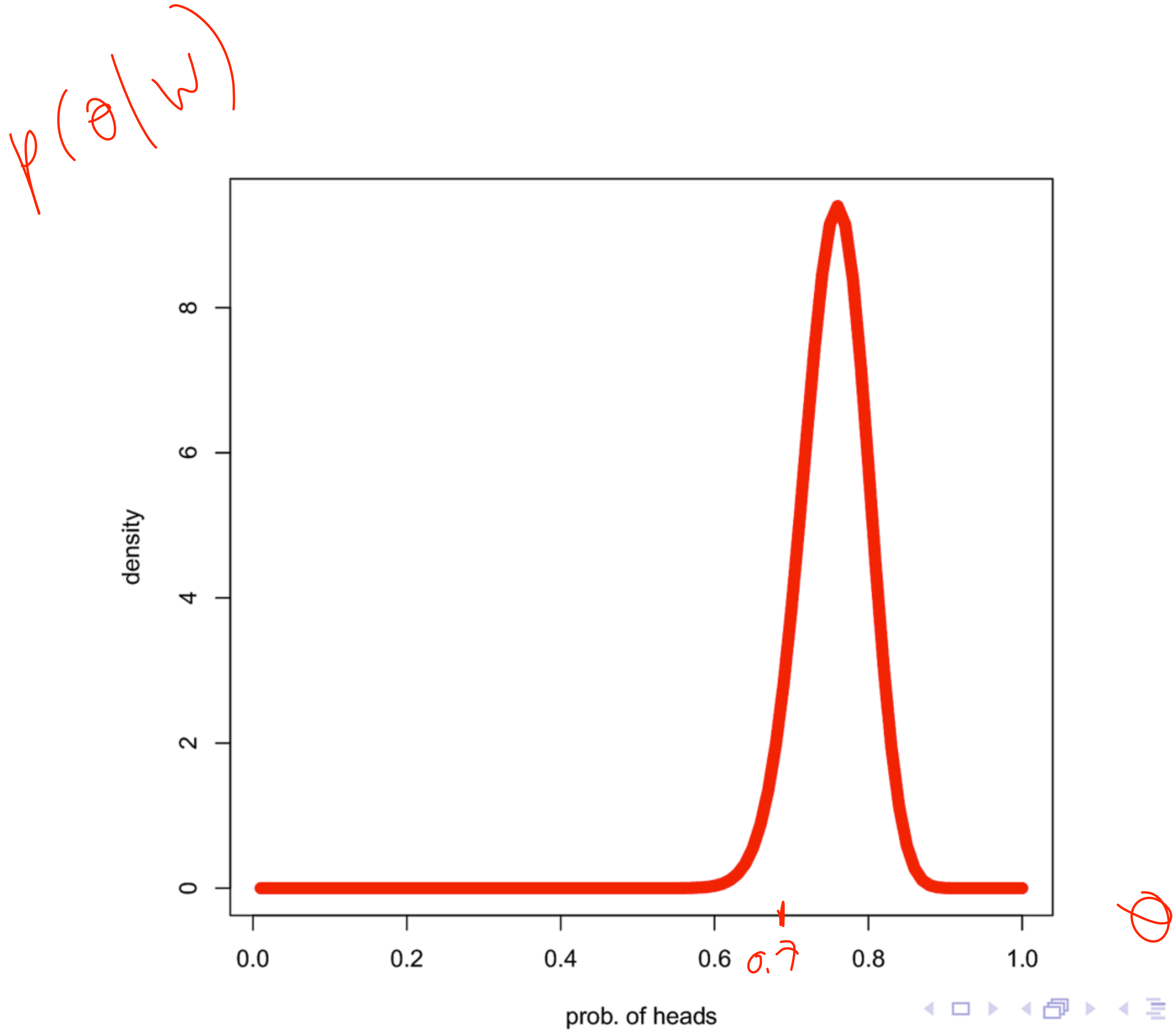
Uniform prior, 0.7 prob. for argh



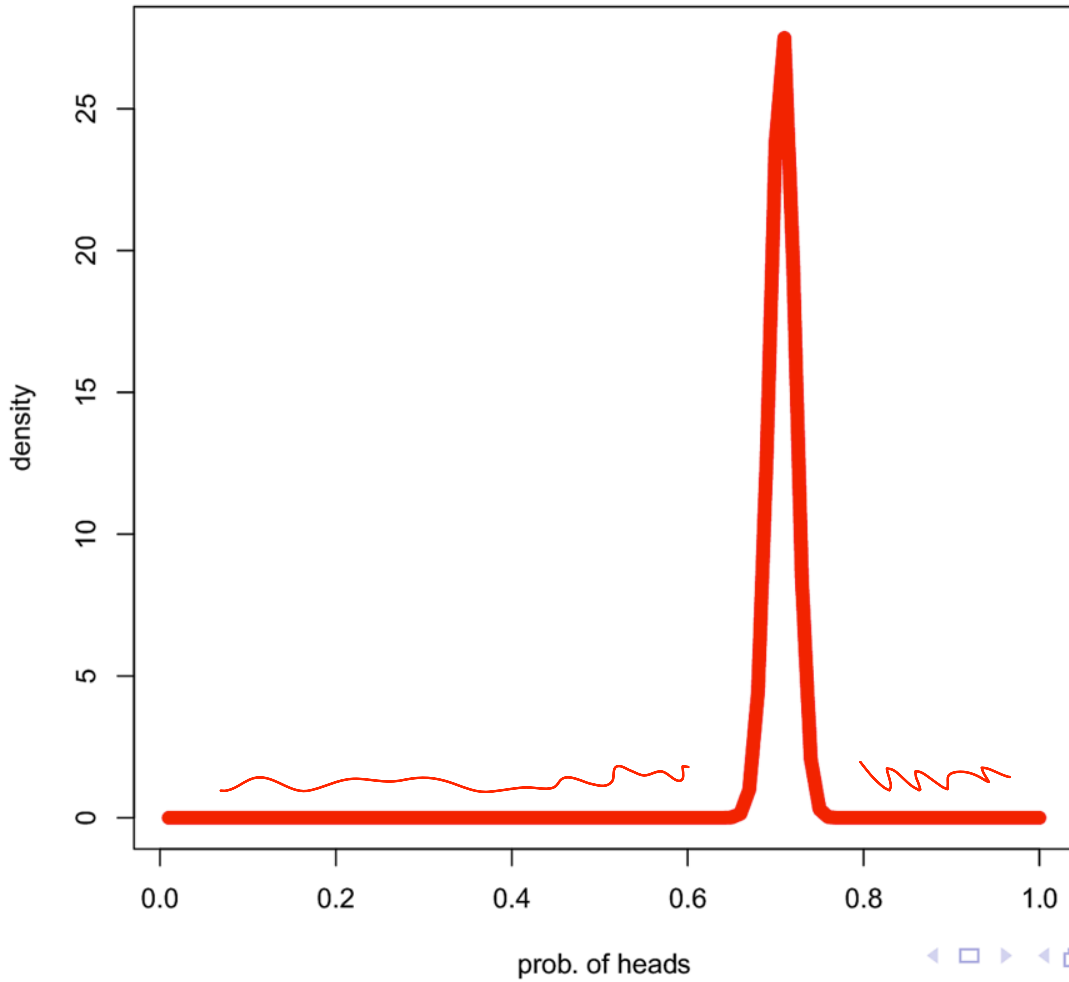
Posterior with 10 datapoints, truth is 0.7 prob. for argh



Posterior with 100 datapoints, truth is 0.7 prob. for argh

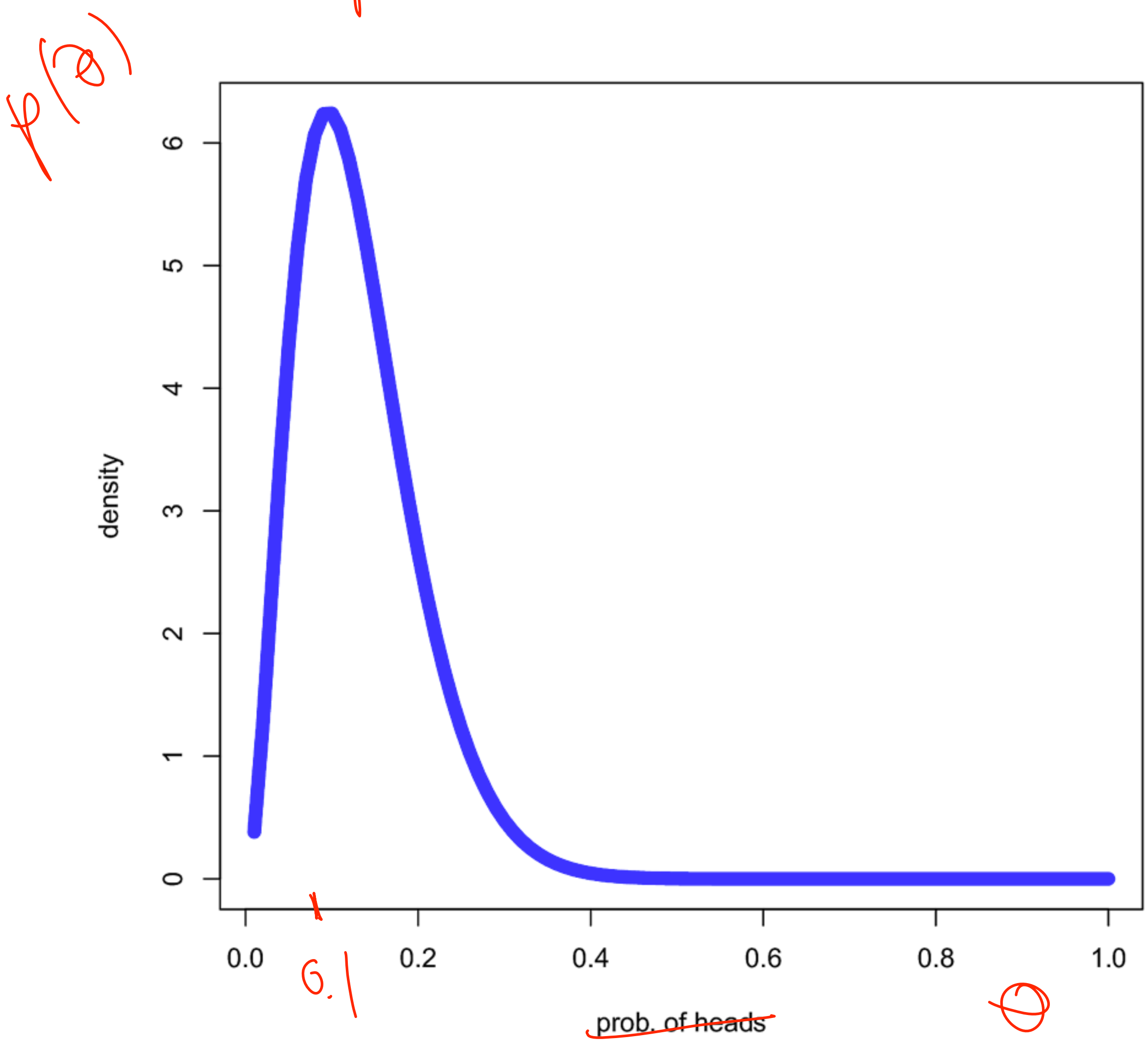


Posterior with 1000 datapoints, truth is 0.7 prob. for argh

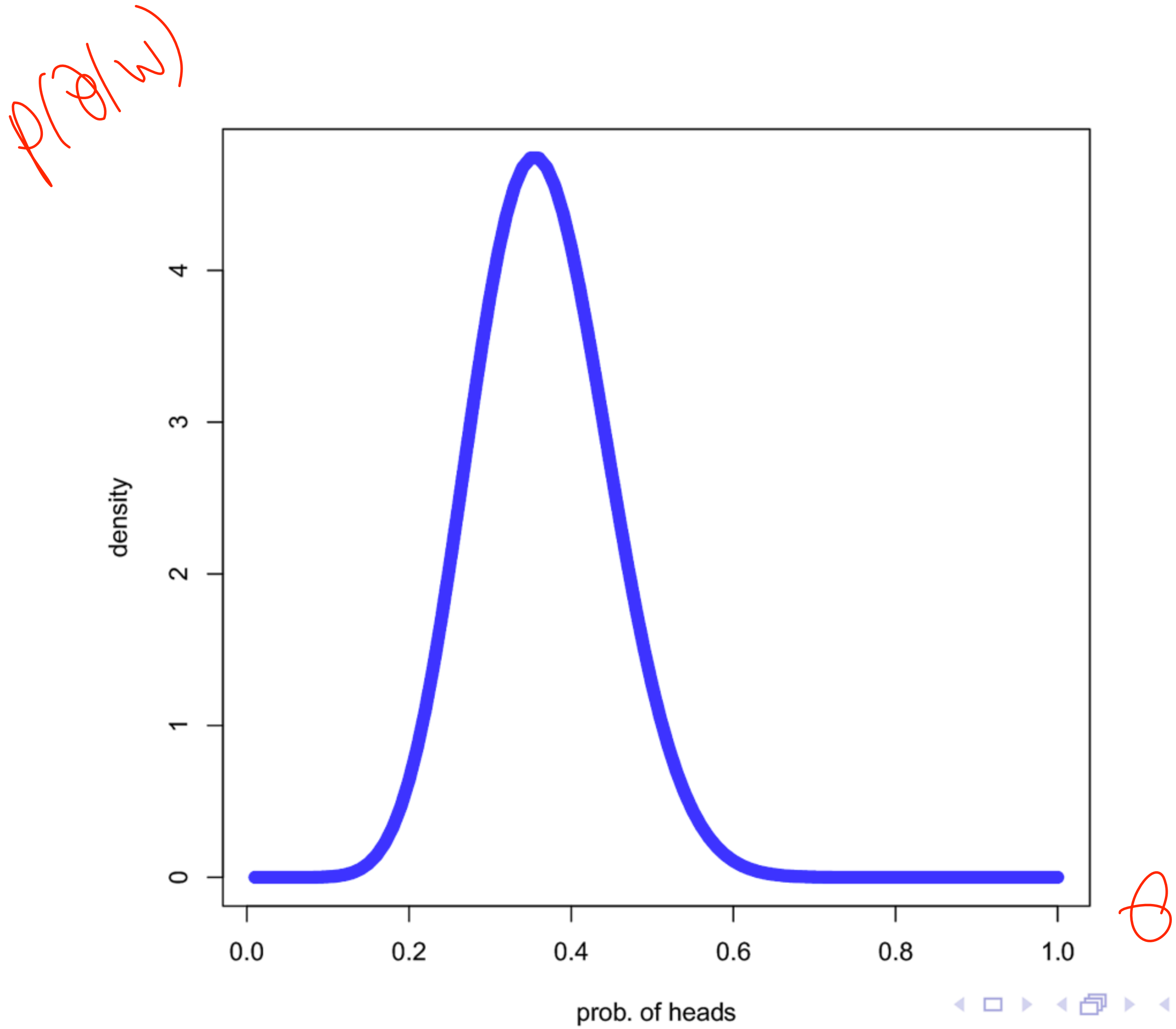


Non-uniform prior, truth is 0.7 prob. for argh

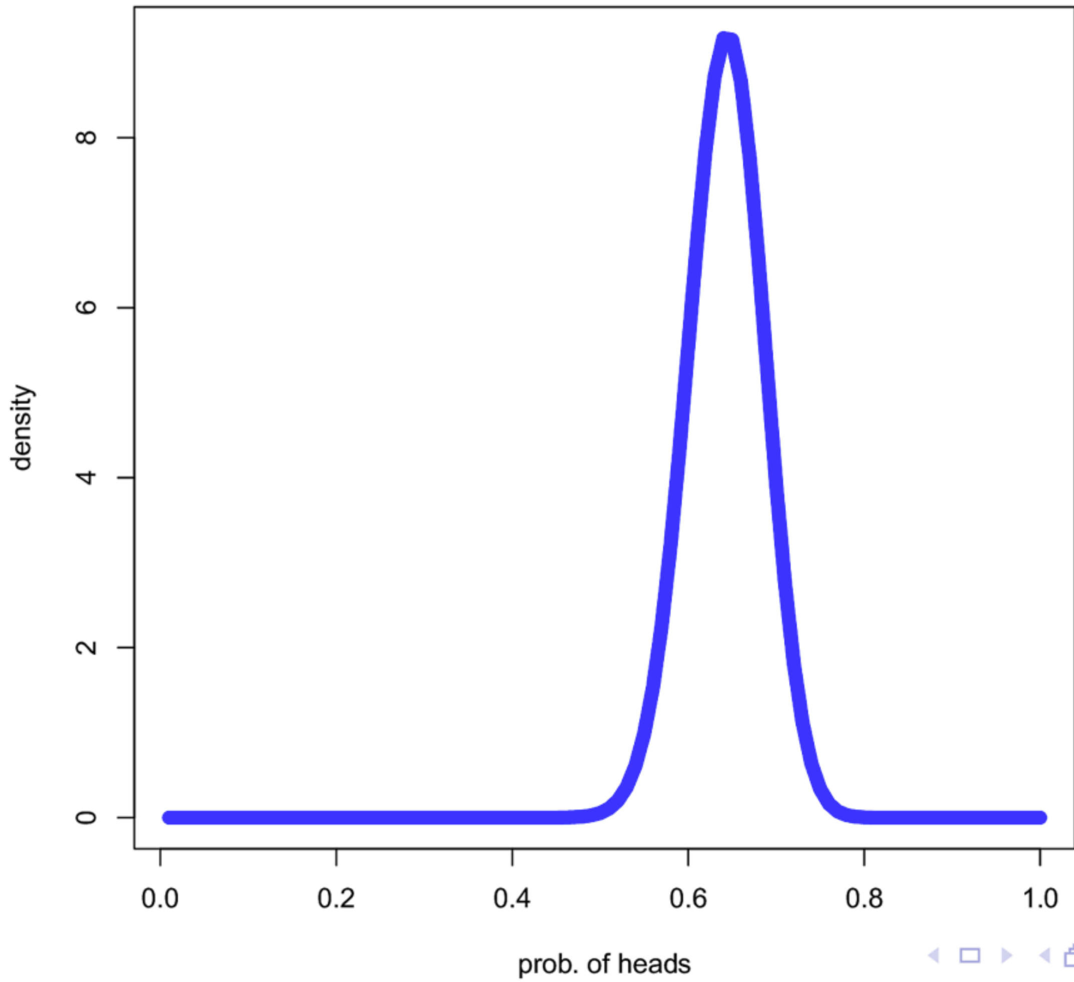
prior



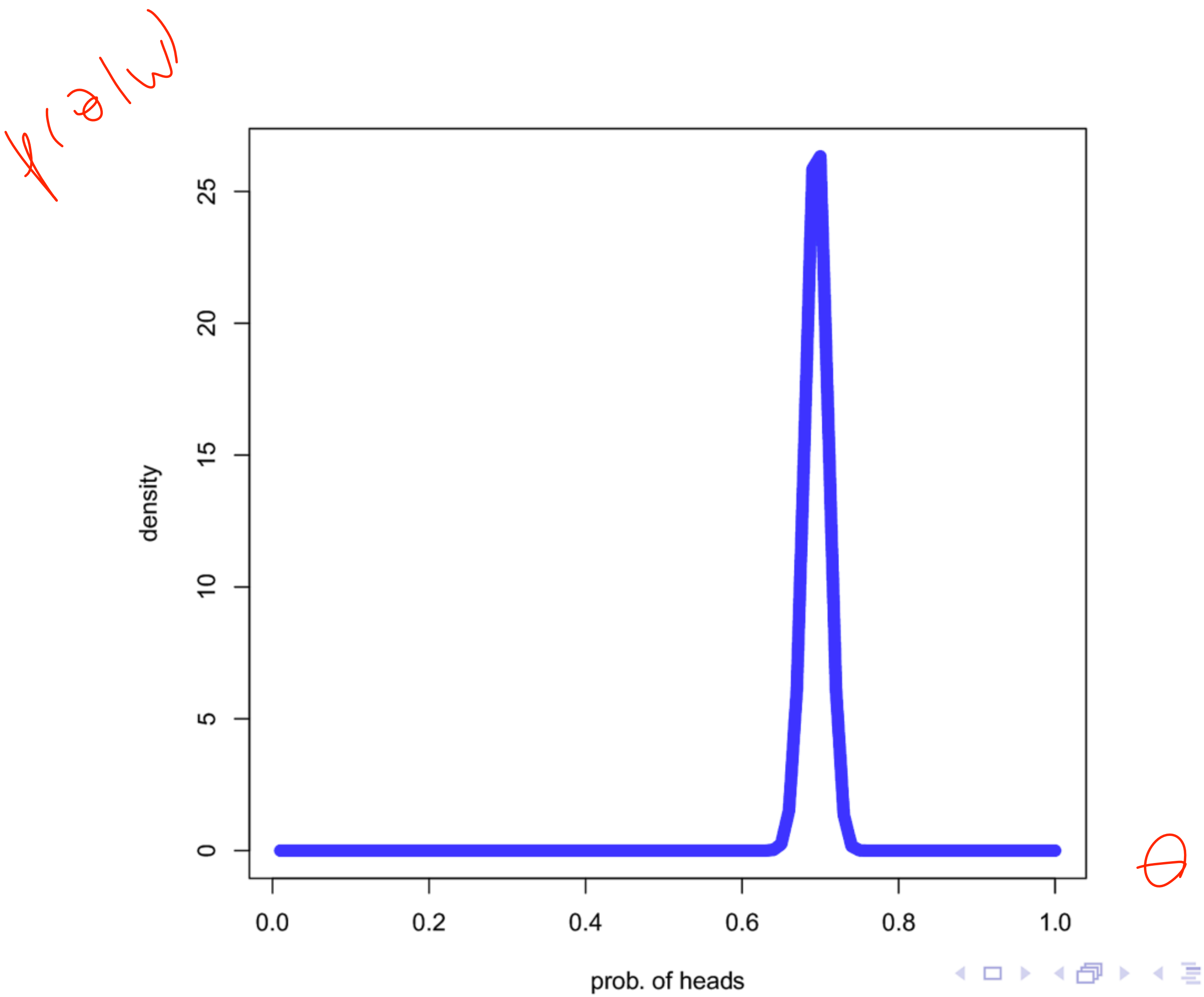
Posterior with 10 datapoints, truth is 0.7 prob. for argh



Posterior with 100 datapoints, truth is 0.7 prob. for argh



Posterior with 1000 datapoints, truth is 0.7 prob. for argh



Priors for binary outcomes

$$p(\theta) \propto \theta^{\alpha}(1-\theta)^{\beta}$$ ← hyperparameter

$$p(w|\theta) = \theta^{I(w)}(1-\theta)^{(1-I(w))}$$

likelihood

What is the posterior?

$D = \{w_1 \ldots w_n\}$

$$p(\theta | w_1 \ldots w_n) = \frac{p(w_1 \ldots w_n | \theta)\, p(\theta)}{p(w_1 \ldots w_n)} = \frac{\left(\prod_{i=1}^{n} p(w_i|\theta)\right) p(\theta)}{p(w_1 \ldots w_n)} = \frac{\prod_{i=1}^{n} \theta^{I(w_i)} (1-\theta)^{1-I(w_i)}}{p(w_1 \ldots w_n)} \times \underline{\theta^{\alpha} \times (1-\theta)^{\beta}}$$

Beta distribution Beta(α, β)

$$= \frac{\theta^{\sum I(w_i)} (1-\theta)^{\sum 1 - I(w_i)} \times \theta^{\alpha} \times (1-\theta)^{\beta}}{p(w_1 \ldots w_n)}$$

$$= \frac{\theta^{\sum I(w_i) + \alpha} (1-\theta)^{n - \sum I(w_i) + \beta}}{p(w_1 \ldots w_n)}$$

$\leftarrow$ doesn't depend on θ

$$\mathrm{Beta}\left(\sum_{i=1}^{n} I(w_i) + \alpha,\; n - \sum_{i=1}^{n} I(w_i) + \beta\right)$$

Maximum a posteriori estimate (MAP)

"Bayesian estimation": find θ^* that maximises the posterior:

$$\theta^*_{MAP} = \arg\max_{\theta} p(\theta | w_1 \ldots w_n) =$$

$$= \arg\max_{\theta} \theta^{a+\alpha} (1-\theta)^{b+\beta}$$

$$= \ldots$$

take logs

$$\theta^*_{MAP} = \frac{a+\alpha}{a+b+\alpha+\beta} = \frac{a+\alpha}{n+(\alpha+\beta)}$$

$$a = \sum I(w_i)$$

$$b = n - \sum I(w_i)$$

MAP and posteriors

In general,

- Priors are especially important when the amount of data is small
- As there is more data, the prior becomes less influential on the posterior
- Under some mild conditions, the posterior is a distribution concentrated around the MLE