Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 3

Administrativia

There is an online forum available now on Piazza. You are able to:

- Ask your peers questions about the material
- Ask your peers general questions about NLP

Last class

Maximum likelihood estimation: $\theta \in \Theta$ $p(w|\theta)$ $\underbrace{w_1 \ldots w_n}_{\text{data}}$

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

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

count of "agh"

$$\Theta = [0,1] \qquad \theta^* = \frac{a}{n}$$

$$\Omega = \{1, \ldots, d\} \qquad \theta_i = \frac{\text{count}(i \text{ in } w_1 \ldots w_n)}{n}$$

$$\Theta \in \mathbb{R}^d$$

Multinomial distribution MLE

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.

Today's class

- The Bayesian paradigm

- If there is time: structure in NLP - or “what is our Ω?”

Some history

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

Bayes' rule

What is Bayes' rule? $p(X=x, Y=y)$

$p(X=x \mid Y=y)$

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

$p(\cancel{X}=y \mid X=x)$

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

Bayes' rule does the same with random variables.

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: $p(\overset{\text{"x"}}{w}, \overset{\text{"y"}}{\theta}) = \underbrace{p(\theta)}_{\text{prior}}\underbrace{p(w|\theta)}_{\text{likelihood}}$

$$\underset{\text{"posterior"}}{p(\theta|w)} = \frac{\overset{\text{"prior"}}{p(\theta)}\,p(w|\theta)}{p(w)}$$

Bayes rule

Posterior inference

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

basic posterior inference

$$p(w) =$$

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

$$\int \frac{p(w \mid \theta)\, p(\theta)}{\boxed{p(w)}}\, d\theta = 1$$

$$\frac{1}{p(w)} \int_\theta p(w \mid \theta)\, p(\theta)\, d\theta = 1$$

$$p(w) = \underbrace{\int_\theta p(w \mid \theta)\, p(\theta)\, d\theta}_{\text{evidence, normalization constant}} = 1$$

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 $\{\text{argh}, \text{blah}\}$.

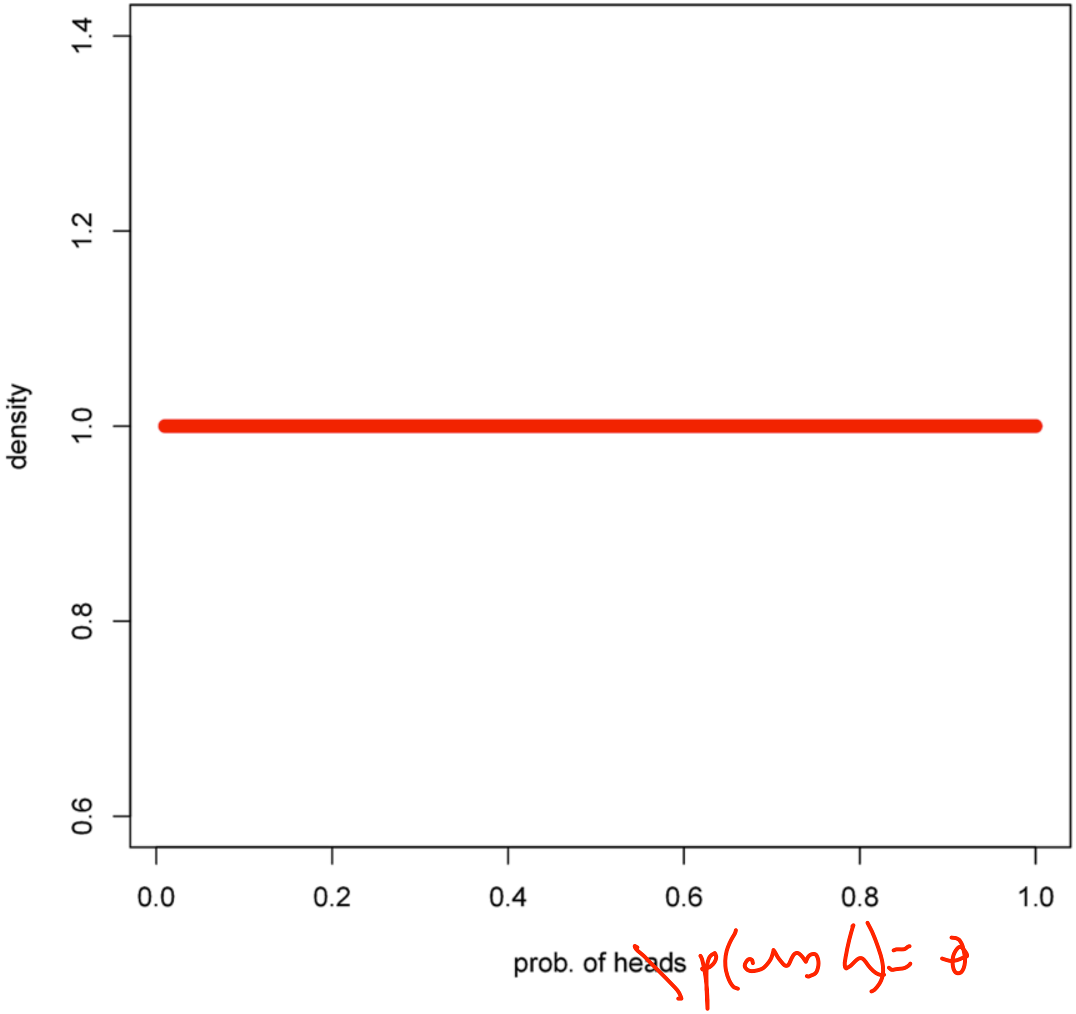
Our Θ is $[0, 1]$.

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

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

Uniform prior, 0.7 prob. for argh

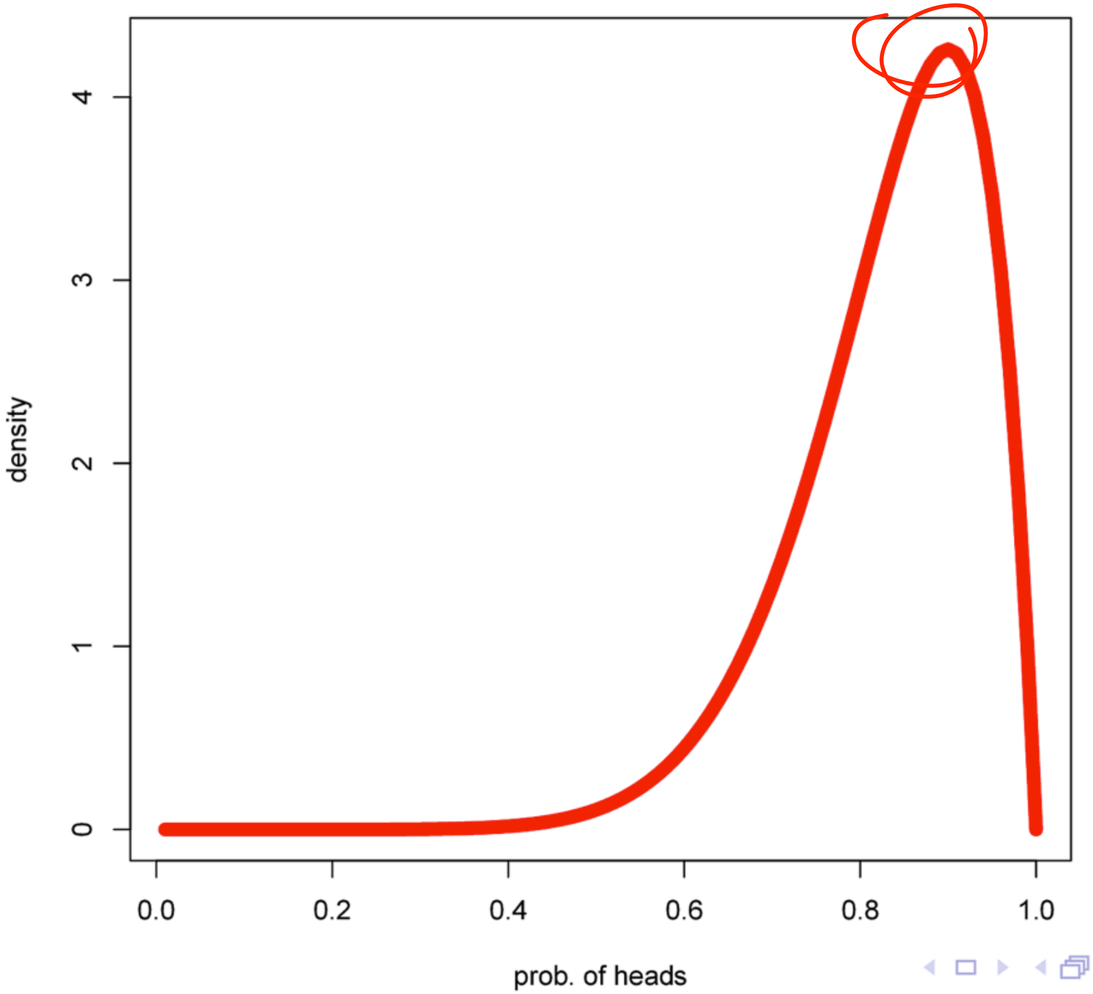
$p(\theta)$



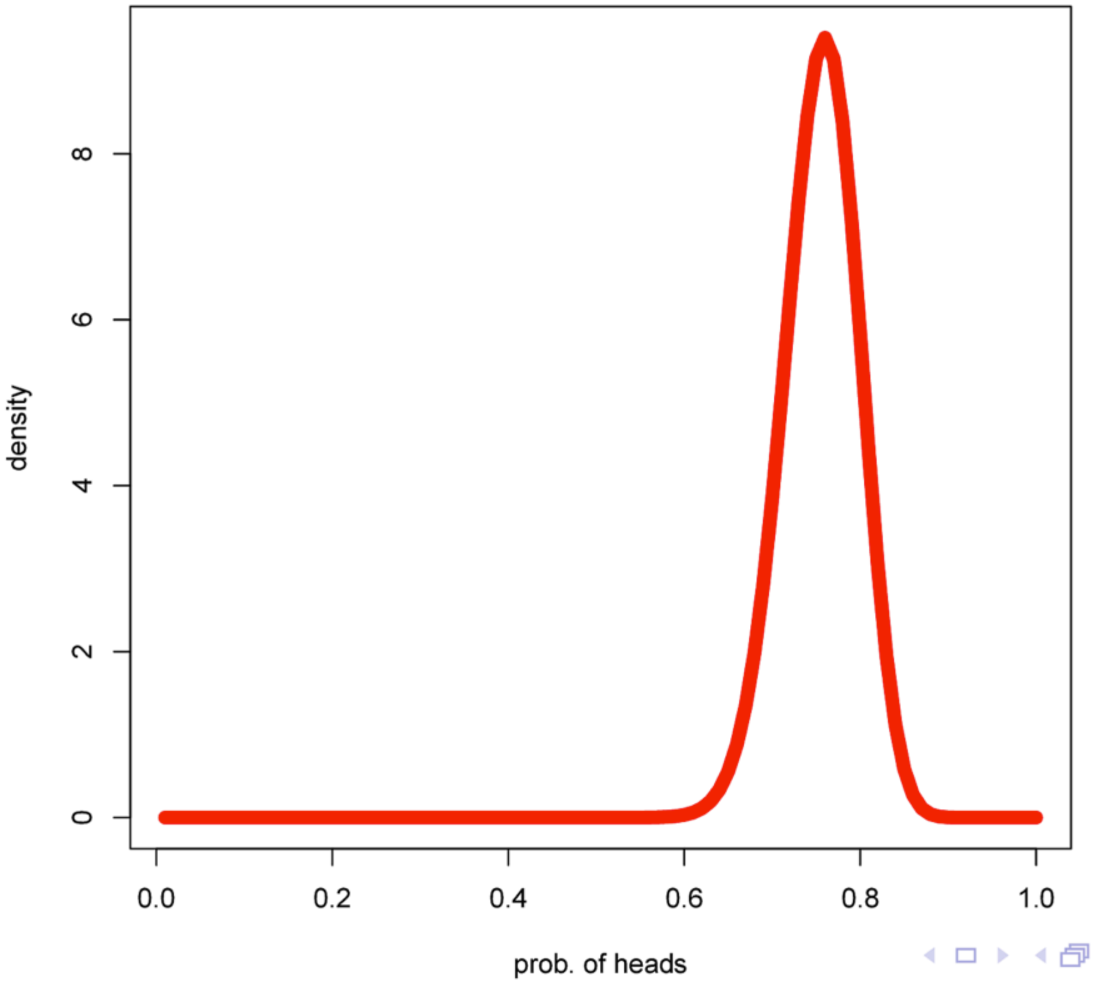
$p(\text{coin } h) = \theta$

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

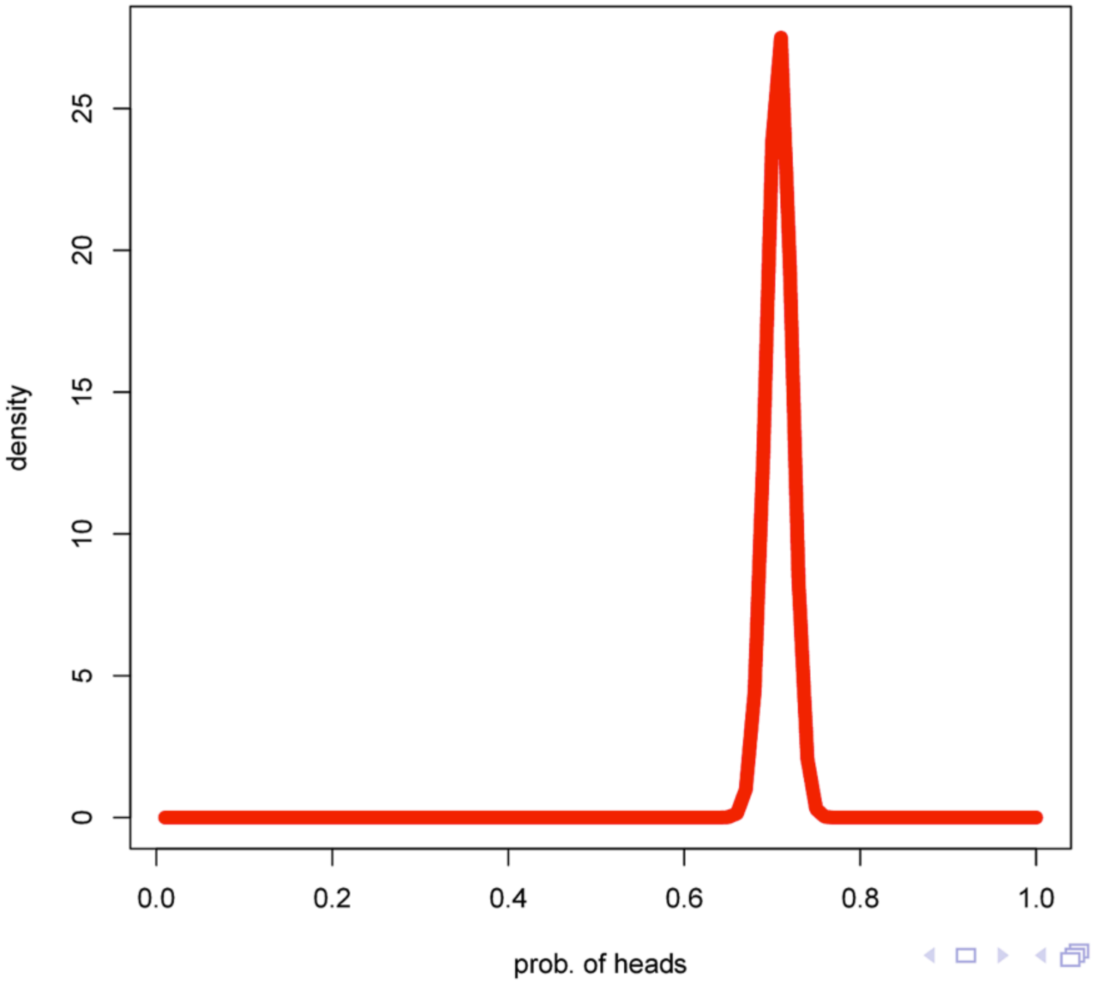
$p(\theta|w_1 \ldots w_{10})$



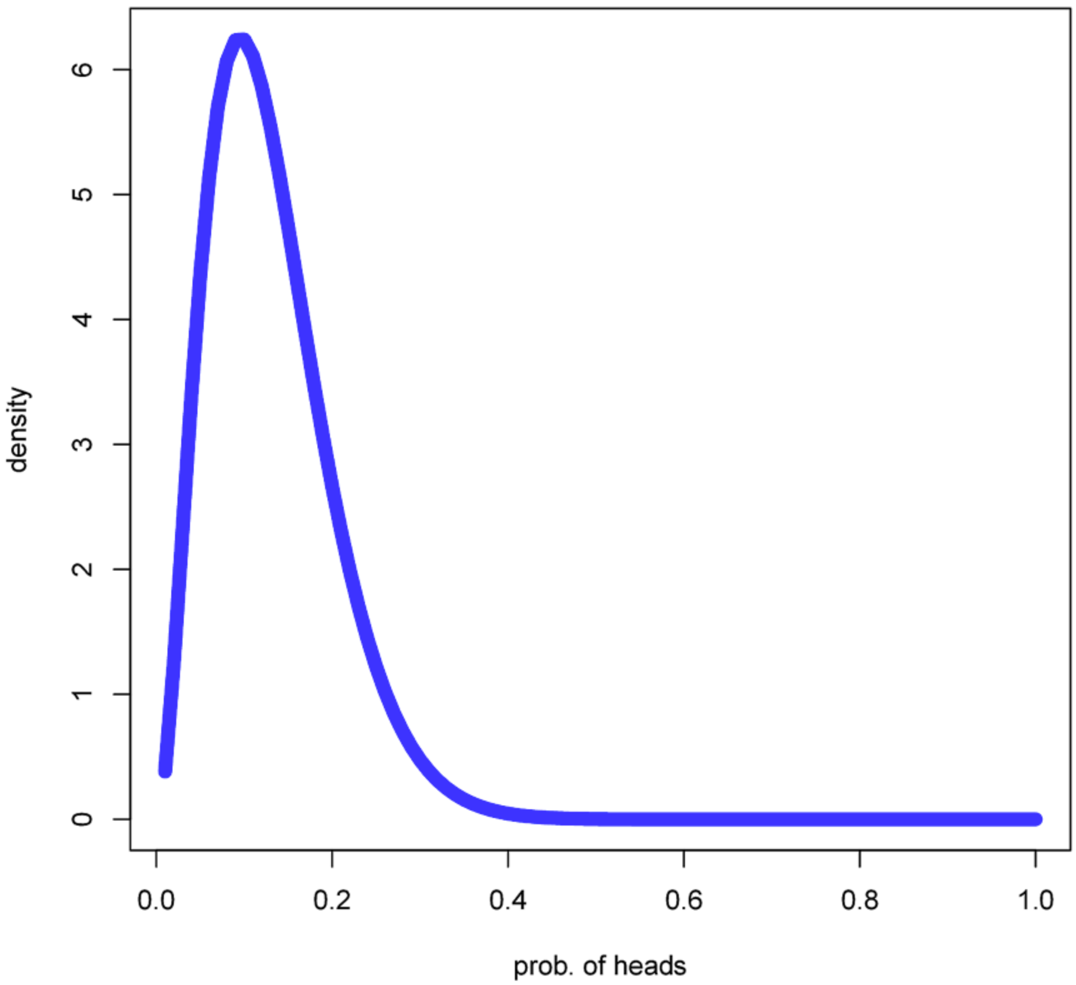
Posterior with 100 datapoints, truth is 0.7 prob. for argh



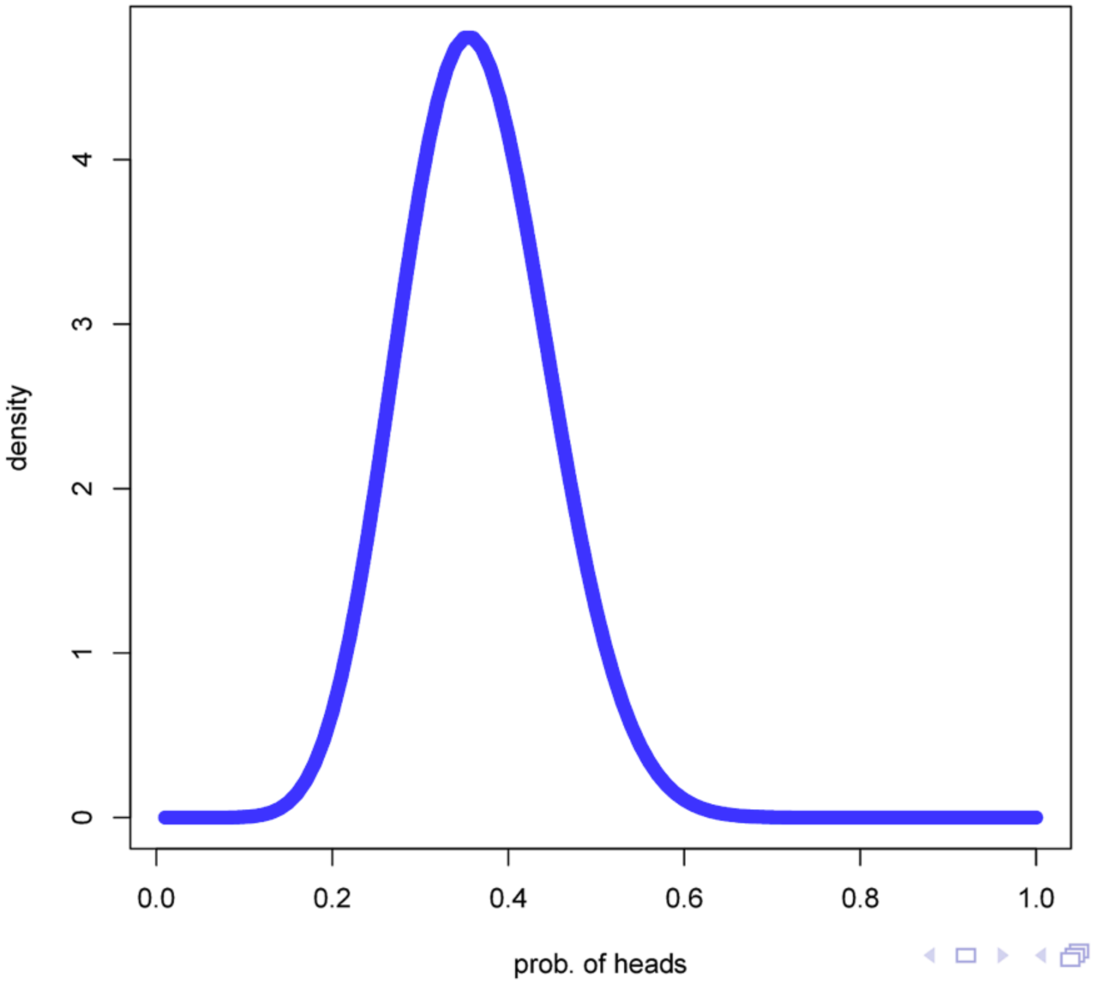
Posterior with 1000 datapoints, truth is 0.7 prob. for argh



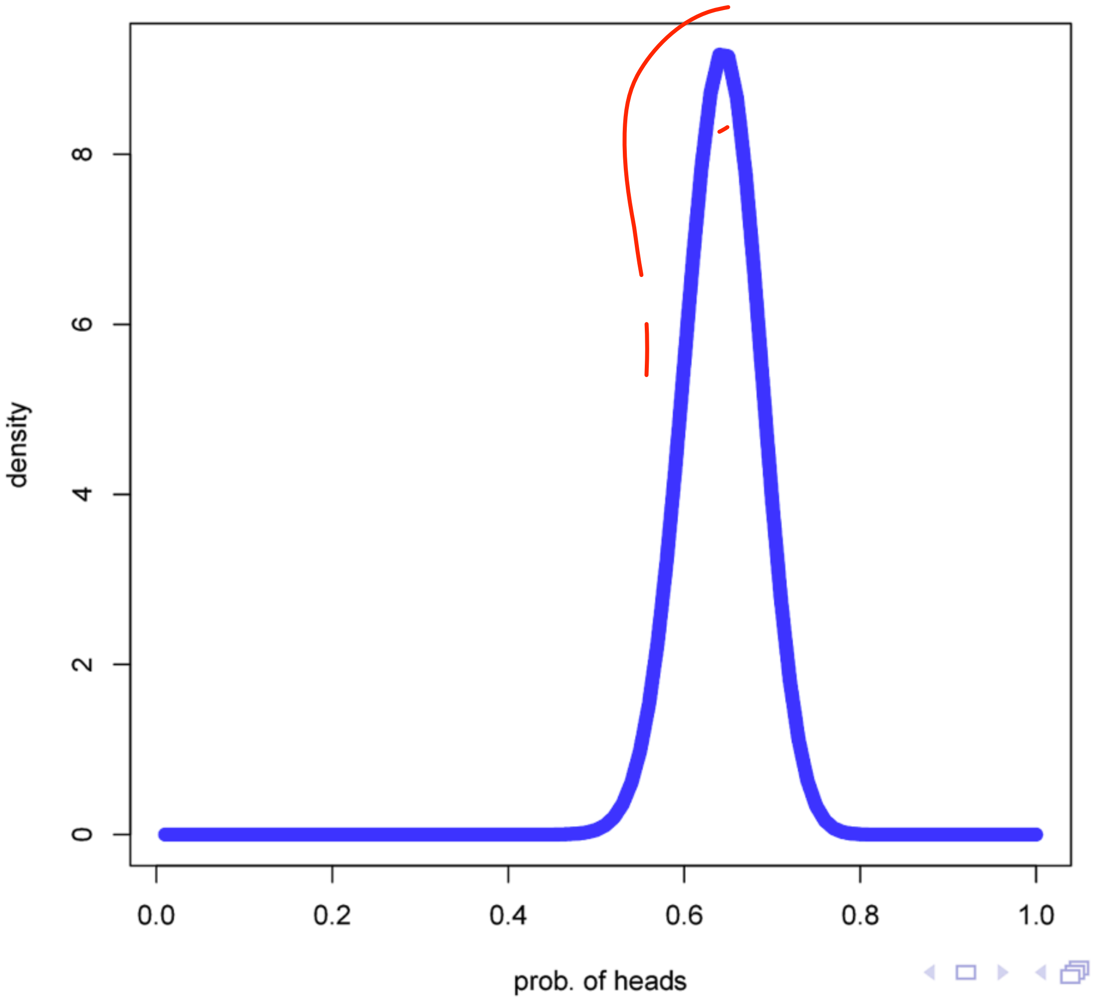
Non-uniform prior, truth is 0.7 prob. for argh



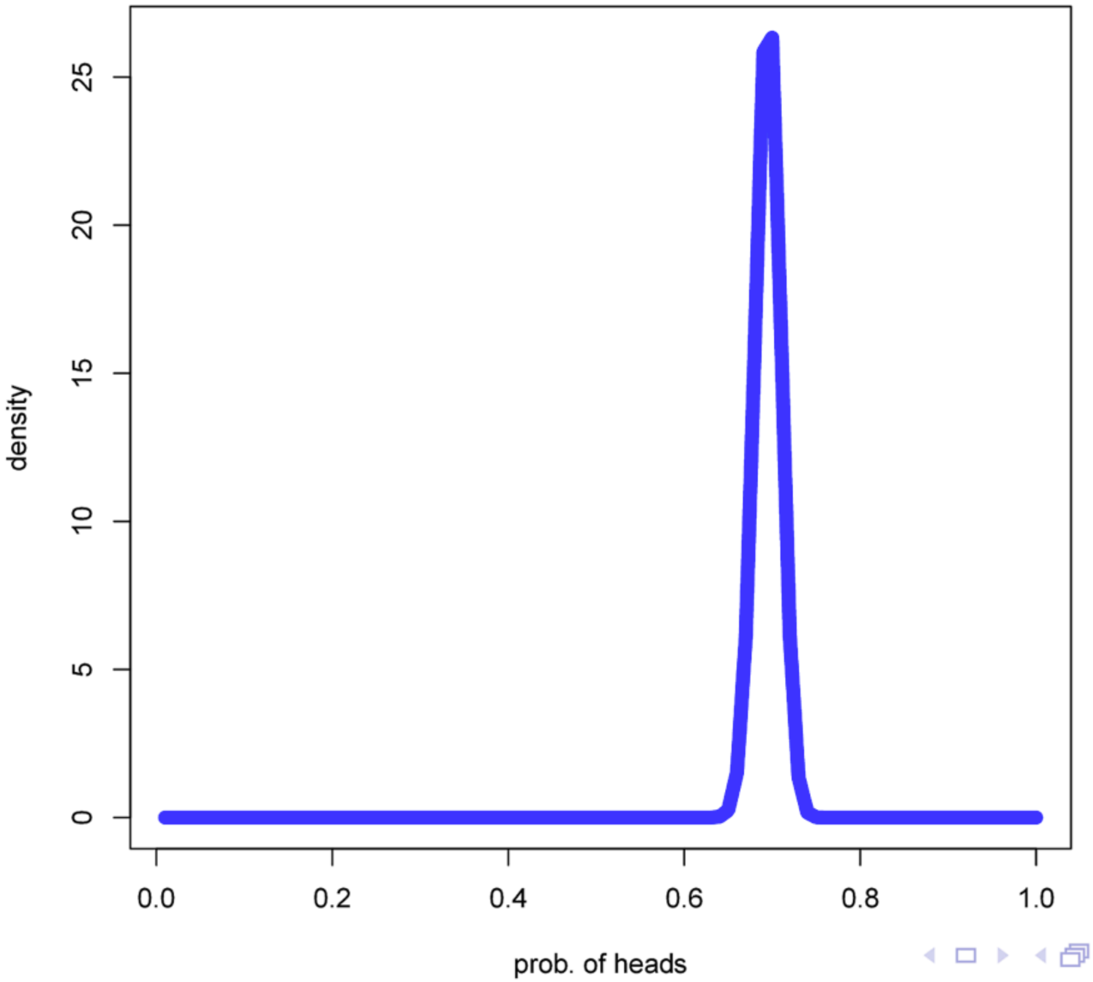
Posterior with 10 datapoints, truth is 0.7 prob. for argh



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

Beta dist

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

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

if $\alpha=\beta=1$ $p(\theta)\propto 1$. which means uniform

What is the posterior? $D=\{w_1 \dots w_n\}$

$$\underline{p(\theta|w_1 \dots w_n)} = \frac{p(w_1 \dots w_n|\theta)\,p(\theta)}{p(w_1 \dots w_n)} = \frac{\left(\prod_{i=1}^{n} p(w_i|\theta)\right)p(\theta)}{p(w_1 \dots w_n)}$$

likelihood

$$= \overbrace{\left(\prod_{i=1}^{n} \theta^{I(w_i)}(1-\theta)^{1-I(w_i)}\right)} \times \theta^{\alpha}(1-\theta)^{\beta} / p(w_1 \dots w_n)$$

$$= \theta^{\sum_{i=1}^{n} I(w_i)}(1-\theta)^{\sum_{i=1}^{n} 1-I(w_i)} \times \theta^{\alpha}(1-\theta)^{\beta} / p(w_1 \dots w_n)$$

$$= \theta^{\alpha+\sum I(w_i)}(1-\theta)^{\beta+\sum 1-I(w_i)} \boxed{/p(w_1 \dots w_n)}$$

Maximum a posteriori estimate (MAP)

“Bayesian estimation”: find θ^* that maximises the posterior:

$$\theta^* = \arg\max_{\theta} p(\theta | w_1 \ldots w_n) = \arg\max_{\theta} \theta^{a+\alpha}(1-\theta)^{b+\beta}$$

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

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

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

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

Next class

- Conjugacy of Bayesian priors to the likelihood
- Structure in NLP