Lecture 18: Probabilistic topic models II: LDA (part 1)

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 18

What we'll learn in this lecture

- Frequentist versus Bayesian thinking
- Prior, posteriors, and conjugacy
- The LDA generative model

Frequentist reasoning

In frequentist reasoning, we produce point estimates of parameters:

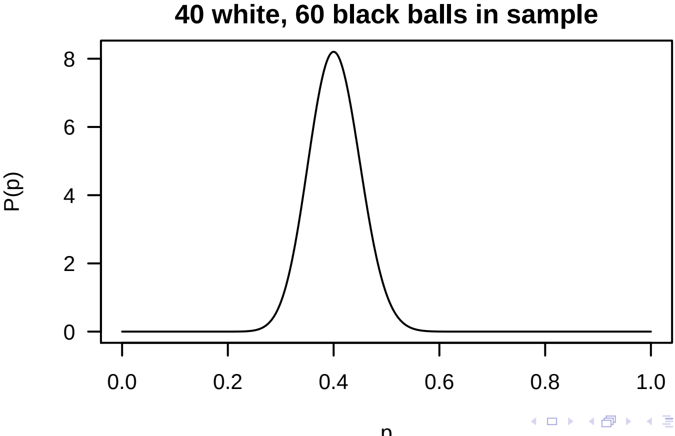
- If we sample 100 balls from a bag of black and white balls, and 40 are white, then:
  - we estimate that 40% of the balls in the bag are white
  - we have 95% confidence that the proportion of balls is between 30.3% and 50.3%
    - (Even the latter statement is more carefully hedged in frequentist reasoning)

Roughly speaking, maximum likelihood estimates live in the frequentist world. (NOTE: all of this discussion is "roughly speaking")

Bayesian reasoning

In Bayesian reasoning, we produce probability distributions over parameters:

If we sample 100 balls from a bag of black and white balls, and 40 are white, then the probability distribution of the proportion p of white balls in the bag looks like:



Bayesian reasoning

- We can do more interesting things with a distribution than a point estimate
- Therefore Bayesian reasoning is more powerful than frequentist reasoning
- However, it requires stronger assumptions
- In particular, it requires us to assume things about the state of the world in the absence of evidence

Bayes' equation

The core Bayesian tool is Bayes' equation:

$$P(a|b) = \frac{P(b|a)}{P(b)} \cdot P(a) \tag{1}$$

We read this as:

- $P(a)$: our *prior* belief about a (a distribution)
- b: the evidence
- $P(b|a)/P(b)$: the probability of seeing the evidence, given our prior belief in the world
- $P(a|b)$: our *posterior* belief about a, given the evidence (a distribution)

Bayes and balls

$$P(a|b) = \frac{P(b|a)}{P(b)} \cdot P(a) \quad (2)$$

In the example of 40 white out of 100 balls:

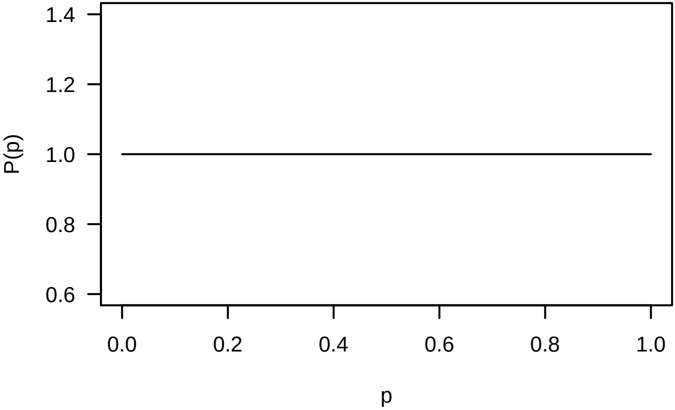
- $P(a)$: our prior belief about the proportion of balls in the bag
- b: the evidence of drawing 100 balls and find 40 white
- $P(b|a)/P(b)$: the probability of drawing 40 white balls given our prior belief
- $P(a|b)$: our posterior belief in the proportion of white balls in the bag

Prior

- Our prior belief, $P(a)$, must be a distribution
- It can't be a single estimate, e.g. 0.5
- because then any outcome except 0.5 is impossible

Prior

Perhaps our prior belief looks like this:

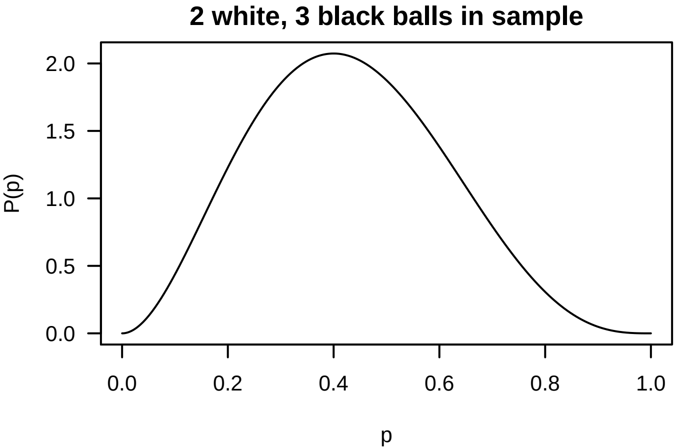


That is, we think any proportion p of white balls is equally likely.
(Note: area under curve is 1, so this is a probability distribution)

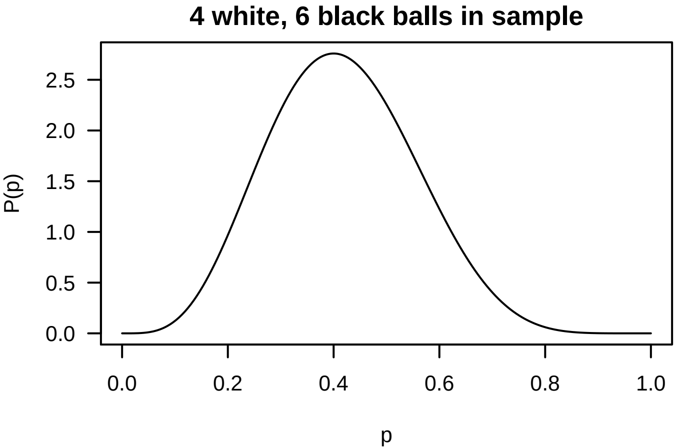
Influence of the prior

- The prior we choose will influence our posterior
- When we have very little evidence, the influence of the prior will be stronger
- As we see more evidence, the influence of the prior will diminish
- . . . and we will put more weight on the evidence
- (This is the way in which a prior "smooths" our belief)

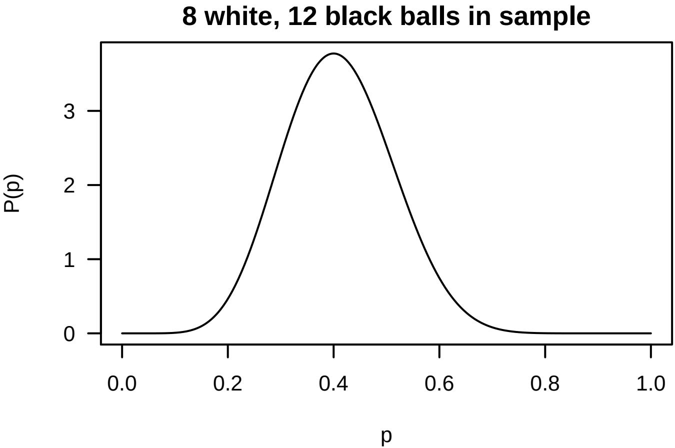
Shifting posterior



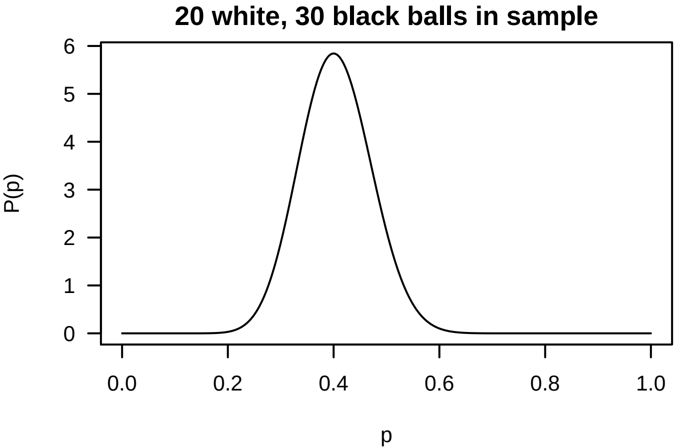
Shifting posterior



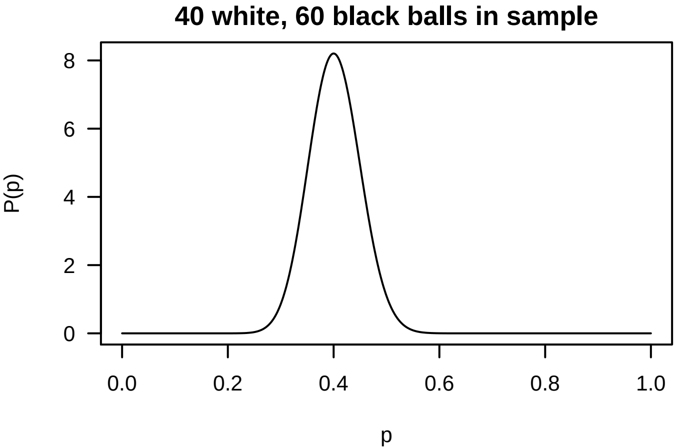
Shifting posterior



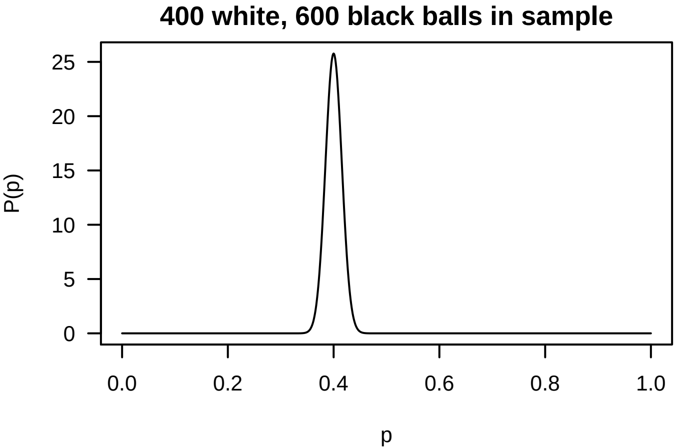
Shifting posterior



Shifting posterior



Shifting posterior



Conjugate prior

$$P(a|b) = \frac{P(b|a)}{P(b)} \cdot P(a) \tag{3}$$

- It is convenient if $P(a)$ and $P(a|b)$ belong to the same family Θ of distributions, albeit with different parameters (say, $\Theta(\alpha)$ and $\Theta(\beta)$)
- The family Φ of $P(b|a)$ will not generally be Θ
- However, we want to choose Θ such that, when $P(a)$ is updated with $P(b|a)$, then $P(a|b)$ is also of family Θ
- When this is the case, we say that Θ is conjugate to Φ (or, equivalently, that $P(a)$ is conjugate prior to $P(b|a)$)

Binomial and beta

$$\begin{aligned} P(b = 1|a) &= p \\ P(b = 0|a) &= 1 - p \end{aligned}$$

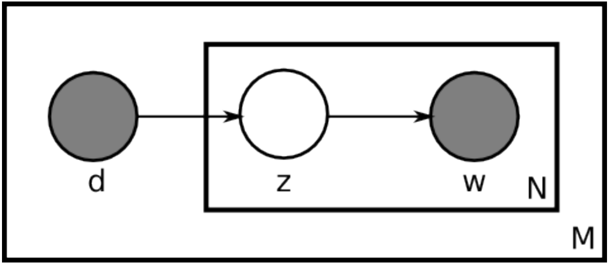
- When b can take only one of two values (white / black, head / tails, true / false), then $P(b|a)$ is *binomial*
- The conjugate prior to the binomial is the *beta* distribution
- Use in this way, the *beta* distribution is a "distribution over distributions" (a meta-distribution)

Multinomial and Dirichlet

$$
\begin{array}{rcl}
P(b = (\text{``}cat''){|}a) & = & q_1 \\
P(b = (\text{``}dog''){|}a) & = & q_2 \\
\dots & & \dots \\
P(b = w_i|a) & = & q_i \\
\dots & & \dots \\
P(b = w_n|a) & = & 1 - \sum_{i}^{n-1} q_i
\end{array}
$$

- When b can take one of $n > 2$ discrete values, the distribution is *multinomial*
- The conjugate prior to the multinomial is the *Dirichlet* distribution

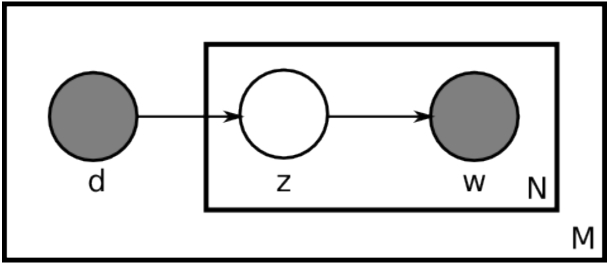
PLSI



$$P(d, w) = P(d) \sum_{z \in \mathcal{Z}} P(w|z = i) P(z = i|d) \tag{4}$$

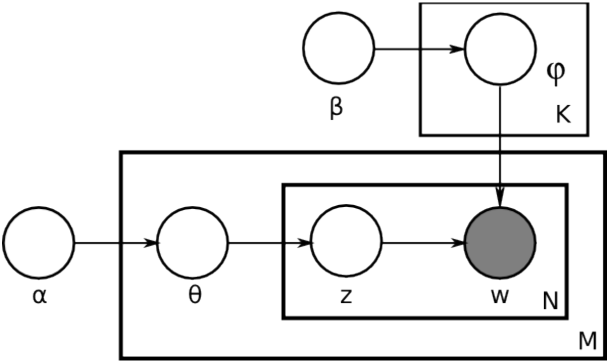
- PLSI is a maximum likelihood method
- Has no principled way of assigning probabilities (e.g. topics) to new document
- Also has no principled way of assigning probabilities to new words

PLSI



- The document d (i.e. document distribution over topics) is an observed variable
- A different distribution over topics is learnt for each of the M documents d
- This requires kM parameters to be found (k is number of topics)
- Leads to over-fitting

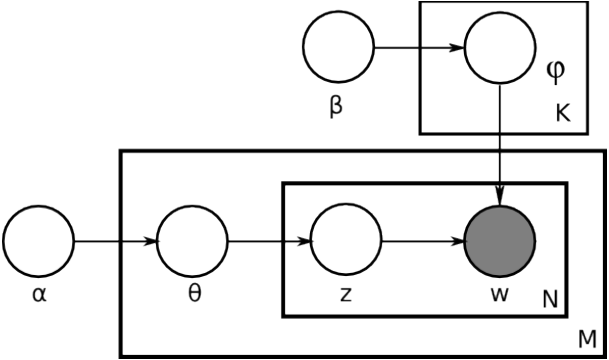
LDA



LDA adds to priors:

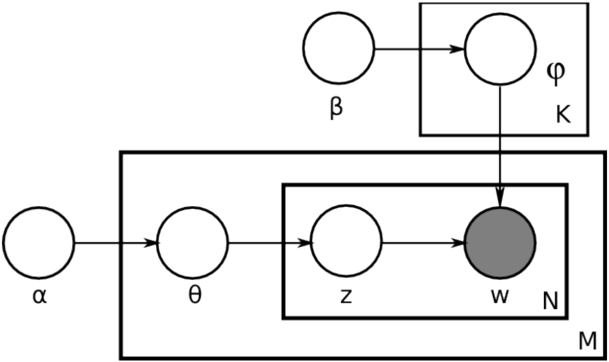
- A prior α to the document distribution over topics
  - Allows us to assign probabilities to new documents
- A prior β over the topic distribution of words
  - Allows us to assign probabilities to new words

LDA



- Each $\Theta \in \{\Theta_1, \ldots, \Theta_M\}$ is a multinomial distribution over topics (given a document)
- Therefore, the prior α to Θ is a Dirichlet distribution
- Each $\Phi \in \{\Phi_1, \ldots, \Phi_K\}$ is a multinomial distribution over a word (given a topic)
- Therefore, the prior β to Θ is also a Dirichlet distribution

LDA



- The Dirichlet priors α and β are not directly observed
- In other words, they are "latent"
- Hence the term "Latent Dirichlet Allocation"

The LDA generative model

The LDA model by which a corpus is formed is as follows:

1. Choose term probabilities for each topic: $\Phi_i \sim \mathcal{D}(\beta)$
2. Choose topic probabilities for each document: $\Theta_d \sim \mathcal{D}(\alpha)$
3. Choose the topic of each token: $z_{dn} \sim \mathcal{M}(\theta_d)$
4. Choose the token: $w_{dn} \sim \mathcal{M}(\phi_{z_{dn}})$

Where:

- $\mathcal{D}$ is a Dirichlet distribution
- $\mathcal{M}$ is a multinomial distribution

Looking back and forward

Back

- Bayesian reasoning produces a posterior distribution over states of the world, based upon a prior and evidence
- The Dirichlet distribution is conjugate prior to the Multinomial
- Latent Dirichlet Allocation “smooths” pLSI by placing Dirichlet priors on:
  - The distribution of topics for a document
  - The distribution of words for a topic

Looking back and forward

Forward

- Next week: finishing the LDA model

Further reading

- Blei, Ng, and Jordan, "Latent Dirichlet Allocation", JMLR, 2003 (the article introducing LDA; note, we are using what they refer to as "smoothed" LDA)
- Crain, Zhou, Yang, and Zha, "Dimensionality Reduction and Topic Modeling", Chapter 5 of Aggarwal and Zhai (ed.), *Mining Text Data*, 2012 (good summary of topic modeling using LSI, pLSI, and LDA).
- Sun, Deng, and Han, "Probabilistic Models for Text Mining", Chapter 8 of Aggarwal and Zhai (ed.), *Mining Text Data*, 2012 (discusses probabilistic models, including the Dirichlet process, in more detail).