Lecture 10 and 11: Text and Discrete Distributions

Machine Learning 4F13, Michaelmas 2015

Zoubin Ghahramani

Department of Engineering
University of Cambridge

http://mlg.eng.cam.ac.uk/teaching/4f13/

Modelling text documents

Here is an article from the Daily Kos (a US political blog) from Feb 16 2014:

```
GOP abortion foes are criminalizing the doctor-patient relationship

"The doctor-patient relationship." For more than 20 years, conservative propagandists
and their Republican allies have used that four-word bludgeon to beat back universal
health care reform.  In 1994, GOP strategist Bill Kristol warned that "the Clinton
Plan is damaging to the quality of American medicine and to the relationship between
the patient and the doctor." Kristol's successful crusade to derail Bill Clinton's
reform effort was greatly aided by future "death panels" fabulist Betsy McCaughey,
who wrongly warned that Americans would even lose the right to see the doctor of
their choice.  Twelve years later, President George W. Bush proclaimed, "Ours is
a party that understands the best health care system is when the doctor-patient
relationship is central to decision-making."

With the victory of Barack Obama in 2008, GOP spinmeister Frank Luntz told
Republicans obstructing the Affordable Care Act in Congress to once again "call
for the 'protection of the personalized doctor-patient relationship.'" And during the
2012 campaign, the GOP platform declared the party would "ensure the doctor-patient
relationship."
...
```

Why would we model text documents?

How could we model this document?

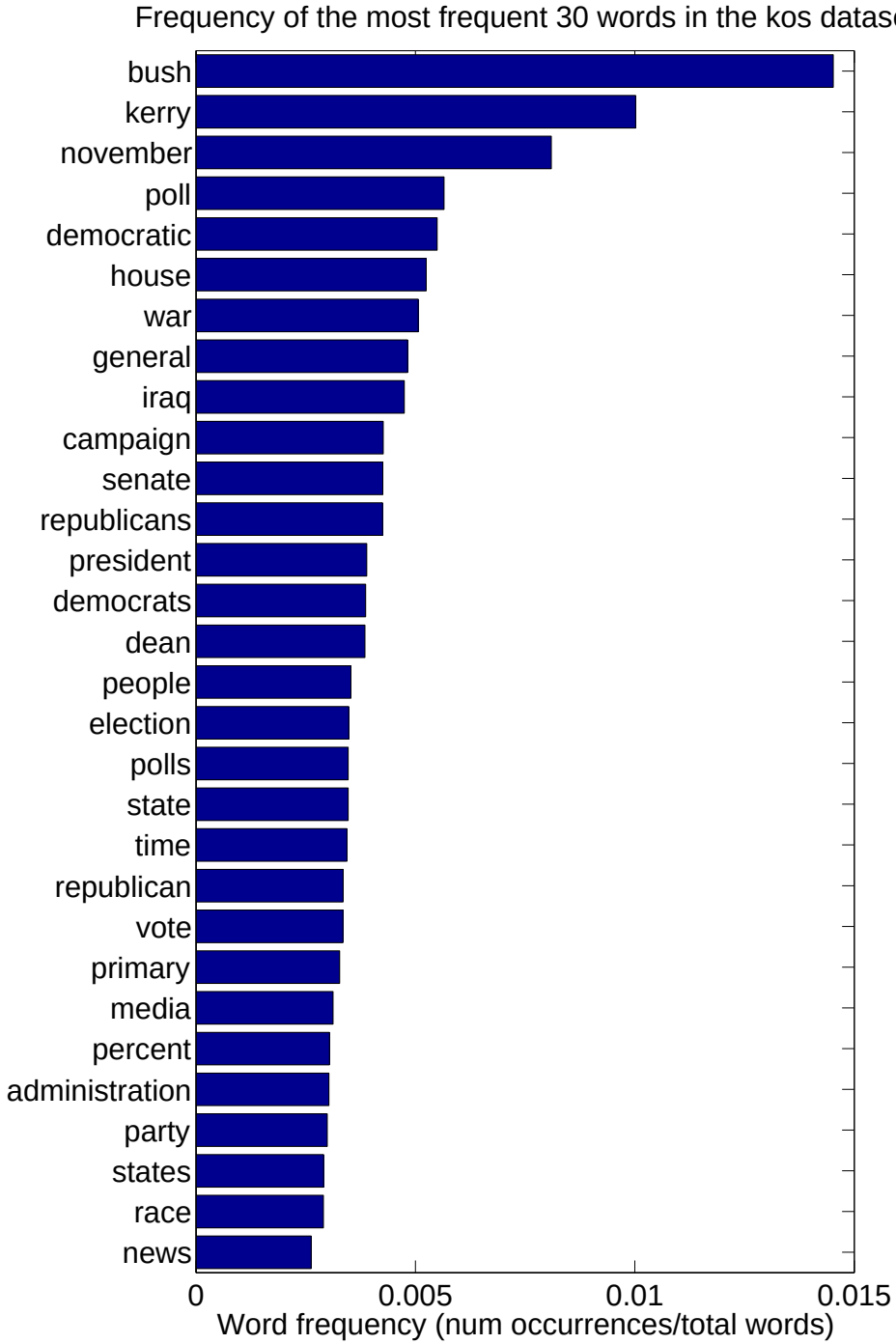
Example: word counts in text

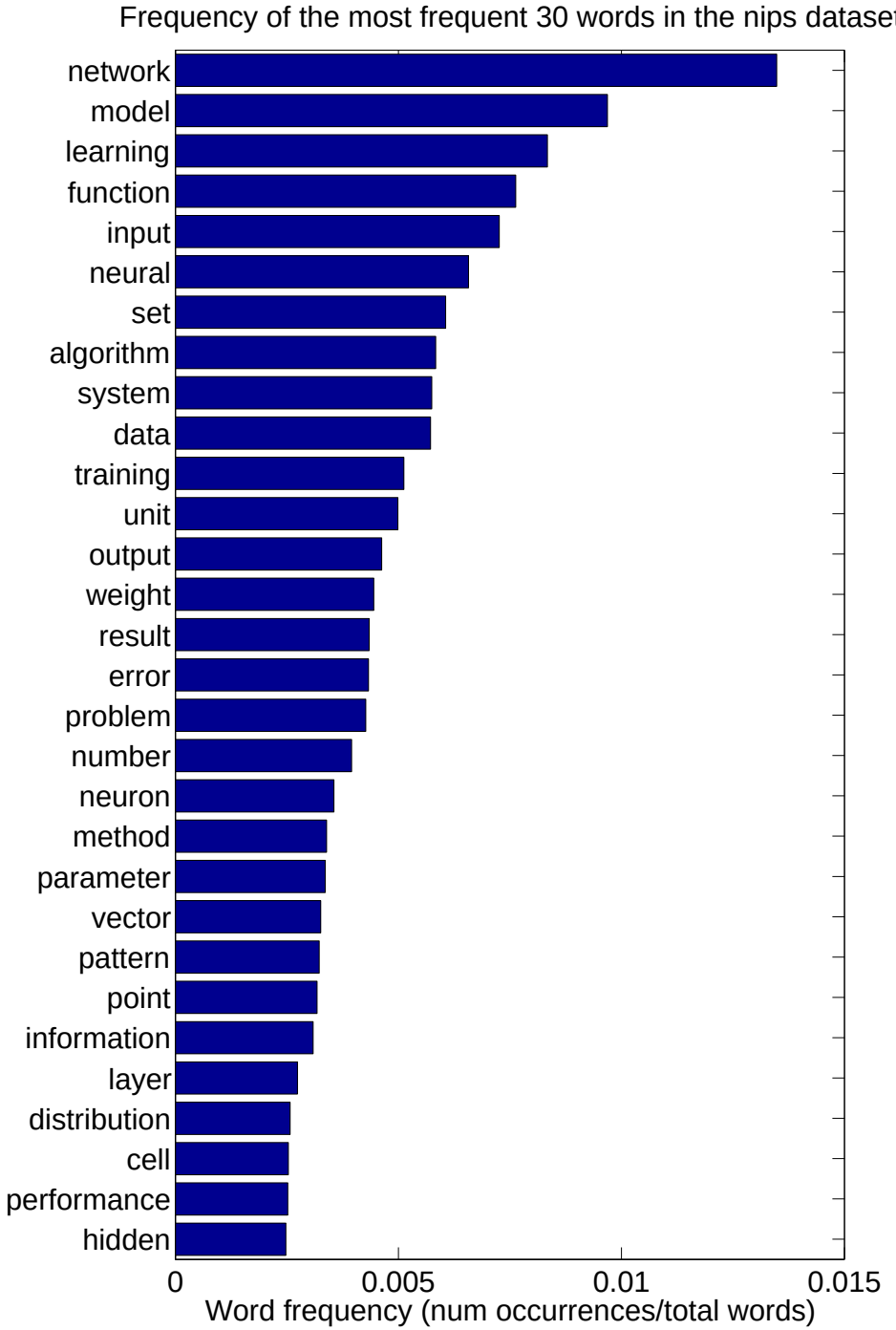
Consider describing a text document by the frequency of occurrence of every distinct word.

The UCI *Bag of Words* dataset from the University of California, Irvine. [1]
For illustration consider two collections of documents from this dataset:

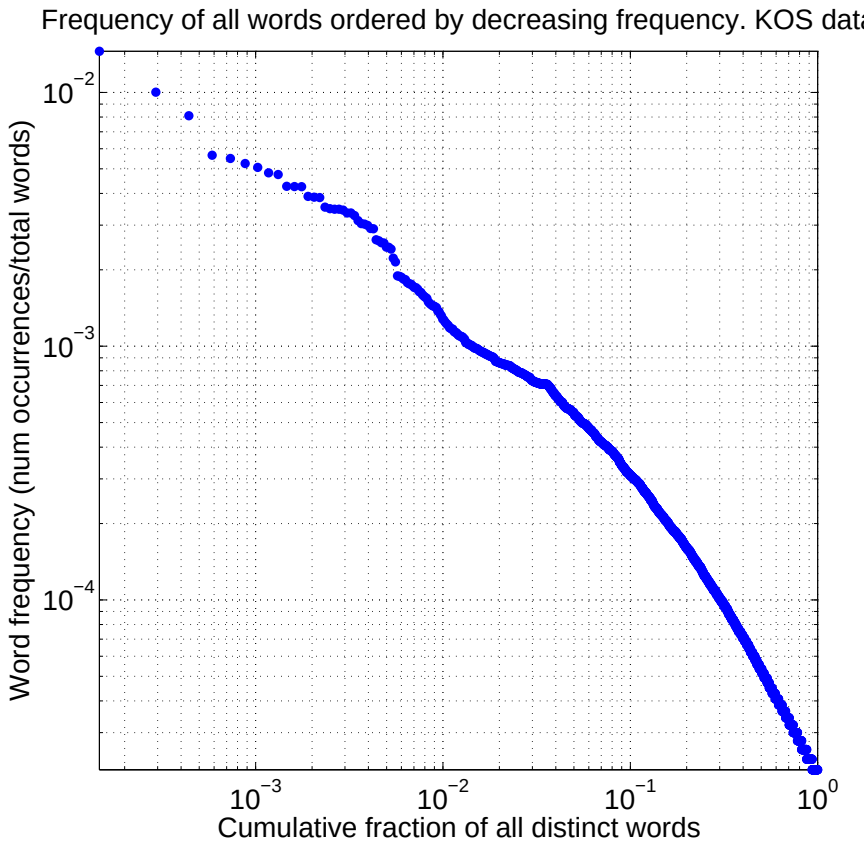
- KOS (political blog — `http://dailykos.com`):
  - $D = 3,430$ documents (blog posts)
  - $n = 353,160$ words
  - $m = 6,906$ *distinct* words
- NIPS (machine learning conference — `http://nips.cc`):
  - $D = 1,500$ documents (conference papers)
  - $n = 746,316$ words
  - $m = 12,375$ *distinct* words

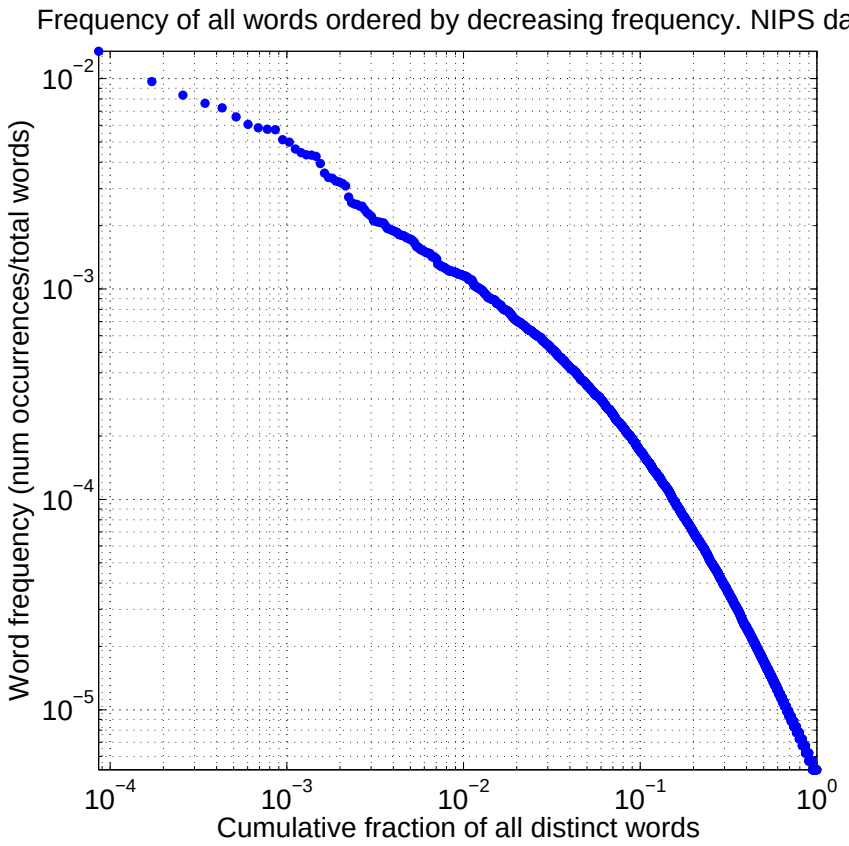
[1] `http://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/`


Frequency of the most frequent 30 words in the kos dataset
bush
kerry
november
poll
democratic
house
war
general
iraq
campaign
senate
republicans
president
democrats
dean
people
election
polls
state
time
republican
vote
primary
media
percent
administration
party
states
race
news
0
0.005
0.01
0.015
Word frequency (num occurrences/total words)


Frequency of the most frequent 30 words in the nips dataset
network
model
learning
function
input
neural
set
algorithm
system
data
training
unit
output
weight
result
error
problem
number
neuron
method
parameter
vector
pattern
point
information
layer
distribution
cell
performance
hidden
0
0.005
0.01
0.015
Word frequency (num occurrences/total words)

Different text collections, similar behaviour





Zipf's law states that *the frequency of any word is inversely proportional to its rank in the frequency table.*

Automatic Categorisation of Documents

Can we make use of the *statistical distribution* of words, to build an automatic document categorisation system?

- The learning system would have to be *unsupervised*
- We don't *a priori* know what categories of documents exist
- It must *automatically discover* the structure of the document collection
- What should it even mean, that a document belongs to a category, or has certain properties?

How can we design such a system?

Coin tossing

- You are presented with a coin: what is the probability of heads?

  What does this question even mean?

- How much are you willing to bet p(head) > 0.5?

  Do you expect this coin to come up heads more often that tails?

  Wait... can you toss the coin a few times, I need data!

- Ok, you observe the following sequence of outcomes (T: tail, H: head):

  H

  This is not enough data!

- Now you observe the outcome of three additional tosses:

  HHTH

  How much are you *now* willing to bet p(head) > 0.5?

The Bernoulli discrete distribution

The *Bernoulli* probability distribution over binary random variables:

- Binary random variable X: outcome x of a single coin toss.
- The two values x can take are
  - $X = 0$ for tail,
  - $X = 1$ for heads.
- Let the probability of heads be $\pi = p(X = 1)$.
  π is the *parameter* of the Bernoulli distribution.
- The probability of tail is $p(X = 0) = 1 - \pi$. We can compactly write

$$p(X = x|\pi) \;=\; p(x|\pi) \;=\; \pi^x(1-\pi)^{1-x}$$

What do we think π is after observing a single heads outcome?

- Maximum likelihood! Maximise $p(H|\pi)$ with respect to π:

$$p(H|\pi) \;=\; p(x = 1|\pi) \;=\; \pi, \qquad \text{argmax}_{\pi\in[0,1]}\, \pi = 1$$

- Ok, so the answer is $\pi = 1$. This coin only generates heads.

Is this reasonable? How much are you willing to bet p(heads)>0.5?

The binomial distribution: counts of binary outcomes

We observe a sequence of tosses rather than a single toss:

HHTH

- The probability of this particular sequence is: $p(\text{HHTH}) = \pi^3(1-\pi)$.
- But so is the probability of THHH, of HTHH and of HHHT.
- We often don't care about the order of the outcomes, only about the *counts*. In our example the probability of 3 heads out of 4 tosses is: $4\pi^3(1-\pi)$.

The *binomial distribution* gives the probability of observing k heads out of n tosses

$$p(k|\pi, n) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

- This assumes n independent tosses from a Bernoulli distribution $p(x|\pi)$.
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient, also known as "n choose k".

Maximum likelihood under a binomial distribution

If we observe k heads out of n tosses, what do we think π is?

We can maximise the likelihood of parameter π given the observed data.

$$p(k|\pi, n) \propto \pi^k (1-\pi)^{n-k}$$

It is convenient to take the logarithm and derivatives with respect to π

$$\log p(k|\pi, n) = k \log \pi + (n-k) \log(1-\pi) + \text{Constant}$$

$$\frac{\partial \log p(k|\pi, n)}{\partial \pi} = \frac{k}{\pi} - \frac{n-k}{1-\pi} = 0 \iff \boxed{\pi = \frac{k}{n}}$$

Is this reasonable?

- For HHTH we get $\pi = 3/4$.
- How much would you bet now that $p(\text{heads}) > 0.5$?

What do you think $p(\pi > 0.5)$*is?*

Wait! This is a probability over ... a probability?

Prior beliefs about coins – before tossing the coin

So you have observed 3 heads out of 4 tosses but are unwilling to bet £100 that $p(\text{heads}) > 0.5$?

(That for example out of 10,000,000 tosses at least 5,000,001 will be heads)

Why?

- You might believe that coins tend to be fair ($\pi \simeq \frac{1}{2}$).
- A finite set of observations *updates your opinion* about π.
- But how to express your opinion about π *before* you see any data?

Pseudo-counts: You think the coin is fair and... you are...

- Not very sure. You act as if you had seen 2 heads and 2 tails before.
- Pretty sure. It is as if you had observed 20 heads and 20 tails before.
- Totally sure. As if you had seen 1000 heads and 1000 tails before.

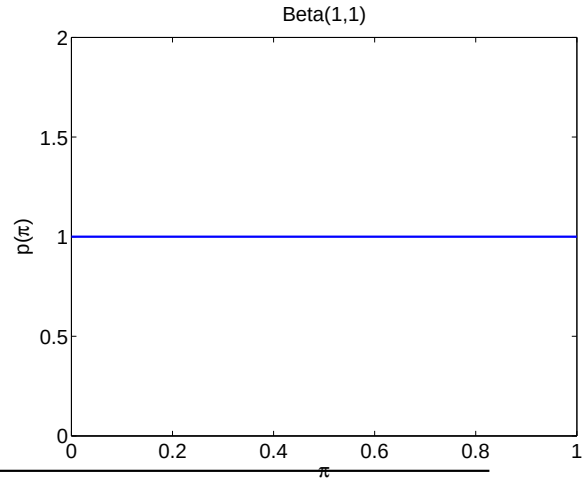
Depending on the strength of your prior assumptions, it takes a different number of actual observations to change your mind.

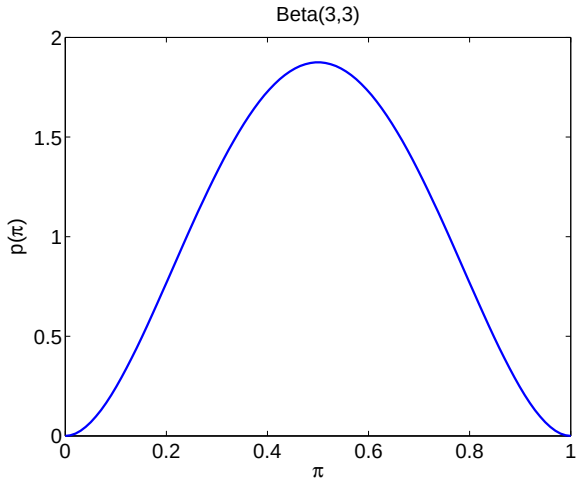
The Beta distribution: distributions on *probabilities*

Continuous probability distribution defined on the interval $[0, 1]$

$$\text{Beta}(\pi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

- $\alpha > 0$ and $\beta > 0$ are the shape *parameters*.
- these parameters correspond to 'one plus the pseudo-counts'.
- $\Gamma(\alpha)$ is an extension of the factorial function[2]. $\Gamma(n) = (n-1)!$ for integer n.
- $B(\alpha,\beta)$ is the beta function, it normalises the Beta distribution.
- The mean is given by $E(\pi) = \frac{\alpha}{\alpha+\beta}$. [Left: $\alpha = \beta = 1$, Right: $\alpha = \beta = 3$]





[2] $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$

Posterior for coin tossing

Imagine we observe a single coin toss and it comes out heads. Our observed data is:

$$\mathcal{D} = \{k = 1\}, \quad \text{where} \quad n = 1.$$

The probability of the observed data given π is the *likelihood*:

$$p(\mathcal{D}|\pi) = \pi$$

We use our *prior* $p(\pi|\alpha, \beta) = \text{Beta}(\pi|\alpha, \beta)$ to get the *posterior* probability:

$$p(\pi|\mathcal{D}) = \frac{p(\pi|\alpha, \beta)p(\mathcal{D}|\pi)}{p(\mathcal{D})} \propto \pi \, \text{Beta}(\pi|\alpha, \beta)$$

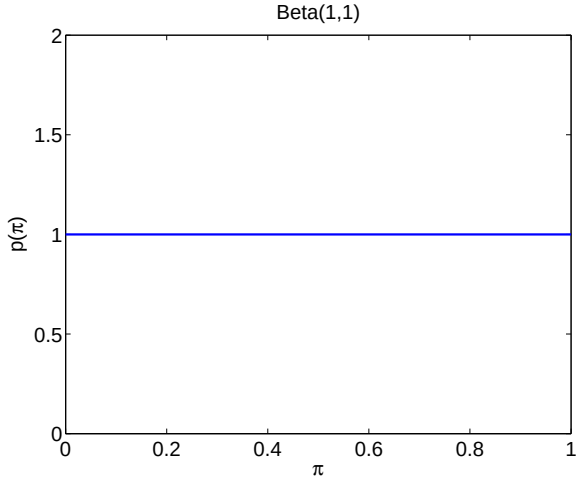
$$\propto \pi \, \pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \propto \text{Beta}(\pi|\alpha+1, \beta)$$

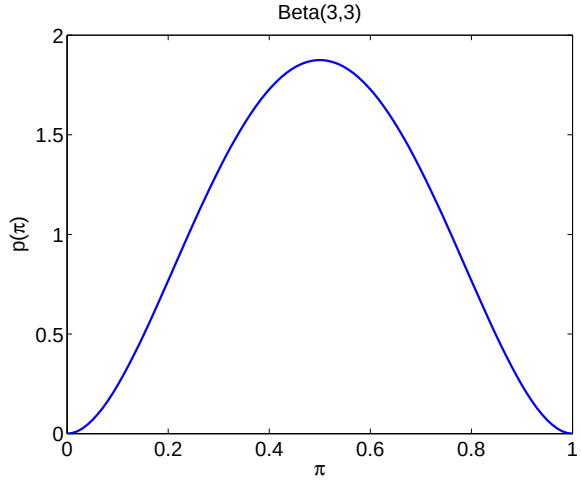
The Beta distribution is a *conjugate* prior to the Bernoulli/binomial distribution:

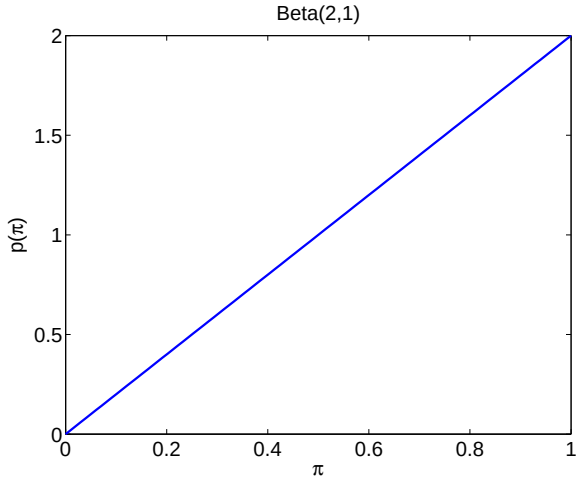
- The resulting posterior is also a Beta distribution.
- The posterior parameters are given by: $\alpha_{\text{posterior}} = \alpha_{\text{prior}} + k$, $\beta_{\text{posterior}} = \beta_{\text{prior}} + (n - k)$

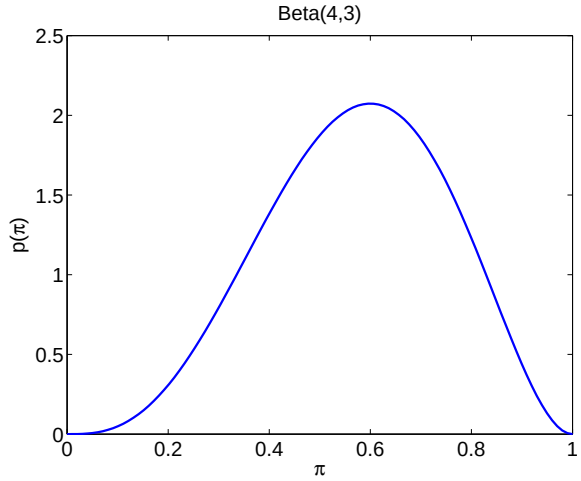
Before and after observing one head

Prior









Posterior

Making predictions

Given some data $\mathcal{D}$, what is the predicted probability of the next toss being heads, $x_{\text{next}} = 1$?

Under the Maximum Likelihood approach we predict using the value of π_{ML} that maximises the likelihood of π given the observed data, $\mathcal{D}$:

$$p(x_{\text{next}} = 1|\pi_{\text{ML}}) = \pi_{\text{ML}}$$

With the Bayesian approach, **average** over all possible parameter settings:

$$p(x_{\text{next}} = 1|\mathcal{D}) = \int p(x = 1|\pi)\, p(\pi|\mathcal{D})\, d\pi$$

The prediction for heads happens to correspond to the mean of the *posterior* distribution. E.g. for $\mathcal{D} = \{(x = 1)\}$:

- Learner A with $\text{Beta}(1, 1)$ predicts $p(x_{\text{next}} = 1|\mathcal{D}) = \frac{2}{3}$
- Learner B with $\text{Beta}(3, 3)$ predicts $p(x_{\text{next}} = 1|\mathcal{D}) = \frac{4}{7}$

Making predictions - other statistics

Given the posterior distribution, we can also answer other questions such as "what is the probability that $\pi > 0.5$ given the observed data?"

$$p(\pi > 0.5|\mathcal{D}) = \int_{0.5}^{1} p(\pi'|\mathcal{D})\, d\pi' = \int_{0.5}^{1} \text{Beta}(\pi'|\alpha', \beta')d\pi'$$

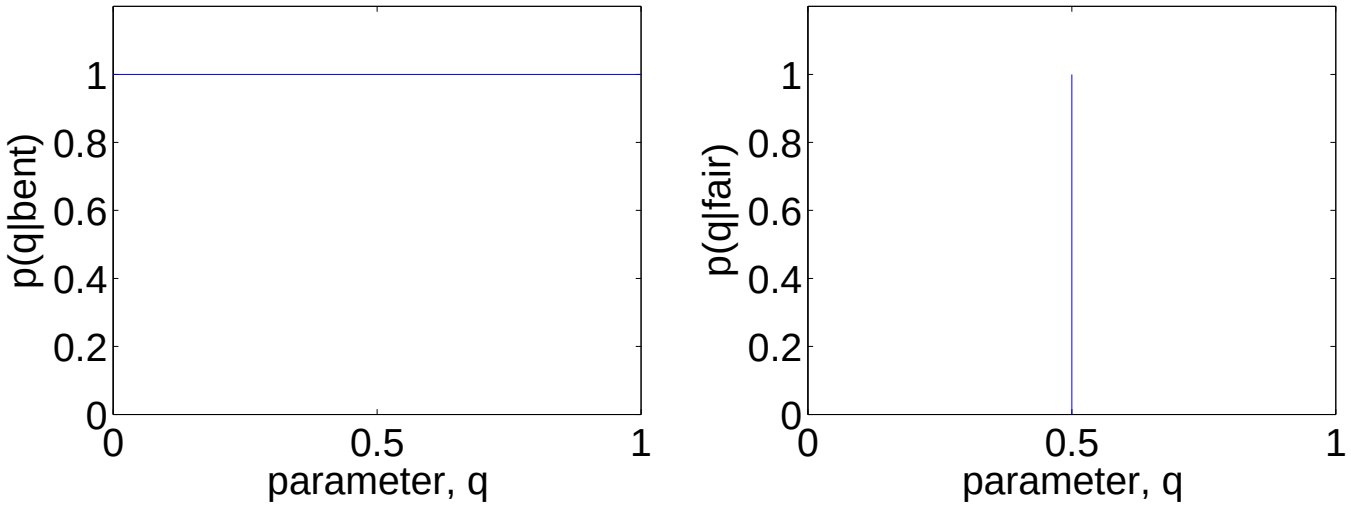
- Learner A with prior $\text{Beta}(1, 1)$ predicts $p(\pi > 0.5|\mathcal{D}) = 0.75$
- Learner B with prior $\text{Beta}(3, 3)$ predicts $p(\pi > 0.5|\mathcal{D}) = 0.66$

Learning about a coin, multiple models (1)

Consider two alternative models of a coin, "fair" and "bent". A priori, we may think that "fair" is more probable, eg:

$$p(\text{fair}) = 0.8, \qquad p(\text{bent}) = 0.2$$

For the bent coin, (a little unrealistically) all parameter values could be equally likely, where the fair coin has a fixed probability:



Learning about a coin, multiple models (2)

We make 10 tosses, and get data $\mathcal{D}$: T H T H T T T T T T

The evidence for the fair model is: $p(\mathcal{D}|\text{fair}) = (1/2)^{10} \simeq 0.001$
and for the bent model:

$$p(\mathcal{D}|\text{bent}) = \int p(\mathcal{D}|\pi, \text{bent}) p(\pi|\text{bent}) \, d\pi = \int \pi^2 (1-\pi)^8 \, d\pi = B(3,9) \simeq 0.002$$

Using priors $p(\text{fair}) = 0.8$, $p(\text{bent}) = 0.2$, the posterior by Bayes rule:

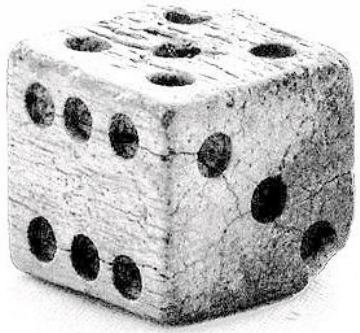
$$p(\text{fair}|\mathcal{D}) \propto 0.0008, \qquad p(\text{bent}|\mathcal{D}) \propto 0.0004,$$

ie, two thirds probability that the coin is fair.

How do we make predictions? By weighting the predictions from each model by their probability. Probability of Head at next toss is:

$$\frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{3}{12} = \frac{5}{12}.$$

The multinomial distribution (1)



Generalisation of the binomial distribution from 2 outcomes to m outcomes.
Useful for random variables that take one of a finite set of possible outcomes.

Throw a die $n = 60$ times, and count the observed (6 possible) outcomes.

Outcome	Count
$X = x_1 = 1$	$k_1 = 12$
$X = x_2 = 2$	$k_2 = 7$
$X = x_3 = 3$	$k_3 = 11$
$X = x_4 = 4$	$k_4 = 8$
$X = x_5 = 5$	$k_5 = 9$
$X = x_6 = 6$	$k_6 = 13$

Note that we have one parameter too many. We don't need to know all the k_i and n, because $\sum_{i=1}^{6} k_i = n$.

The multinomial distribution (2)

Consider a discrete random variable X that can take one of m values $x_1, \ldots, x_m$.

Out of n independent trials, let k_i be the number of times $X = x_i$ was observed. It follows that $\sum_{i=1}^{m} k_i = n$.

Denote by π_i the probability that $X = x_i$, with $\sum_{i=1}^{m} \pi_i = 1$.

The probability of observing a vector of occurrences $\mathbf{k} = [k_1, \ldots, k_m]^\top$ is given by the *multinomial distribution* parametrised by $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$:

$$p(\mathbf{k}|\boldsymbol{\pi}, n) = p(k_1, \ldots, k_m|\pi_1, \ldots, \pi_m, n) = \frac{n!}{k_1!k_2!\ldots k_m!} \prod_{i=1} \pi_i^{k_i}$$

- Note that we can write $p(\mathbf{k}|\boldsymbol{\pi})$ since n is redundant.
- The multinomial coefficient $\frac{n!}{k_1!k_2!\ldots k_m!}$ is a generalisation of $\binom{n}{k}$.

The discrete or *categorical distribution* is the generalisation of the Bernoulli to m outcomes, and the special case of the multinomial with one trial:

$$p(X = x_i|\boldsymbol{\pi}) = \pi_i$$

Example: word counts in text

Consider describing a text document by the frequency of occurrence of every distinct word.

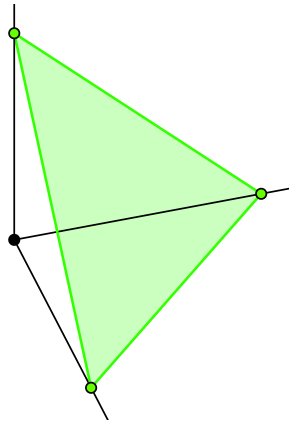
The UCI *Bag of Words* dataset from the University of California, Irvine. [3]

[3] http://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/

Priors on multinomials: The Dirichlet distribution

The Dirichlet distribution is to the categorical/multinomial what the Beta is to the Bernoulli/binomial.
It is a generalisation of the Beta defined on the $m-1$ dimensional simplex.

- Consider the vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$, with $\sum_{i=1}^m \pi_i = 1$ and $\pi_i \in (0,1) \ \ \forall i$.
- Vector $\boldsymbol{\pi}$ lives in the open standard $m-1$ simplex.
- $\boldsymbol{\pi}$ could for example be the parameter vector of a multinomial. [Figure on the right $m = 3$.]

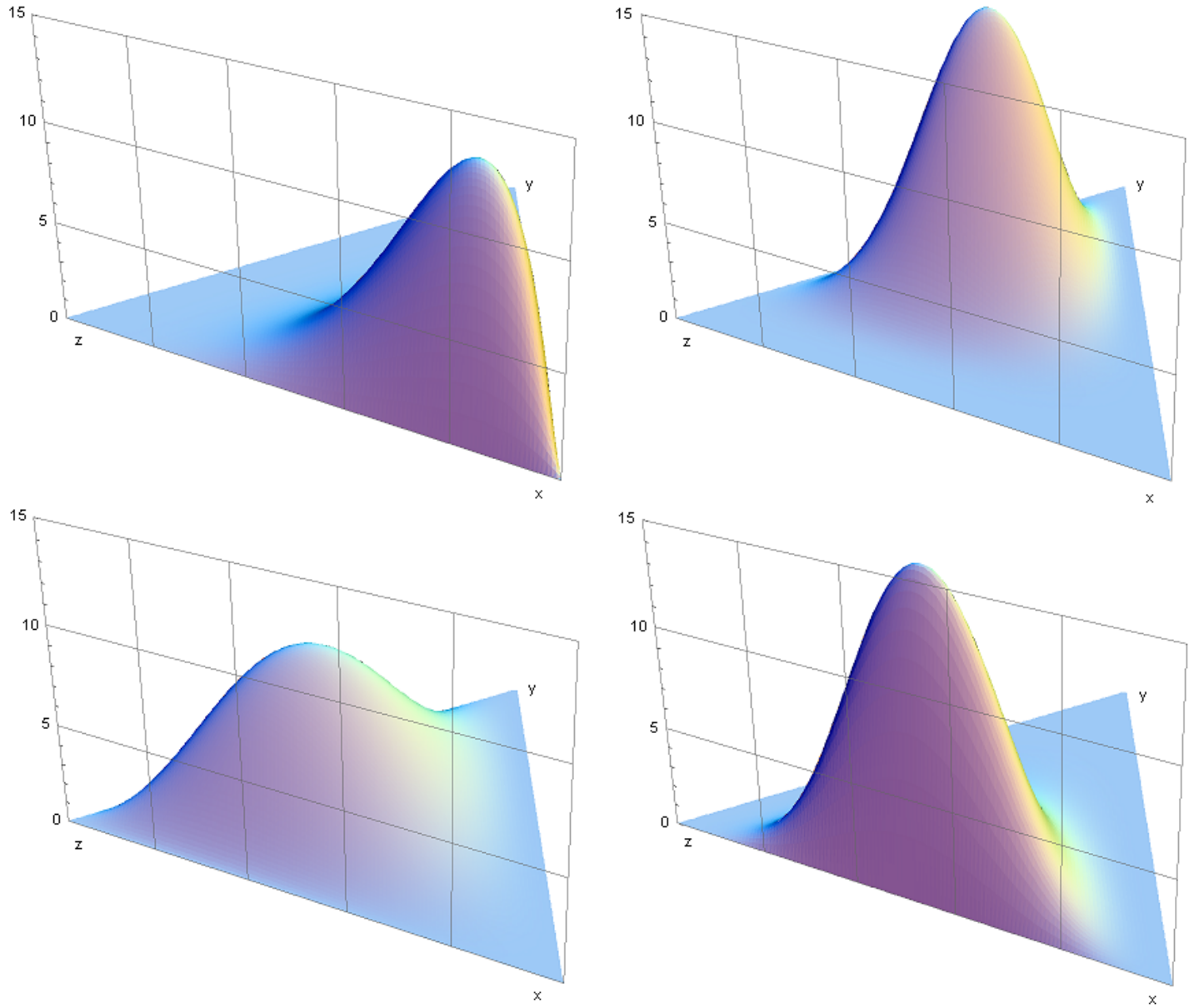


The Dirichlet distribution is given by

$$\text{Dir}(\boldsymbol{\pi}|\alpha_1, \ldots, \alpha_m) = \frac{\Gamma(\sum_{i=1}^m \alpha_i)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m \pi_i^{\alpha_i - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^m \pi_i^{\alpha_i - 1}$$

- $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m]^\top$ are the shape parameters.
- $B(\boldsymbol{\alpha})$ is the multivariate beta function.
- $E(\pi_j) = \frac{\alpha_j}{\sum_{i=1}^m \alpha_i}$ is the mean for the j-th element.

Dirichlet Distributions from Wikipedia

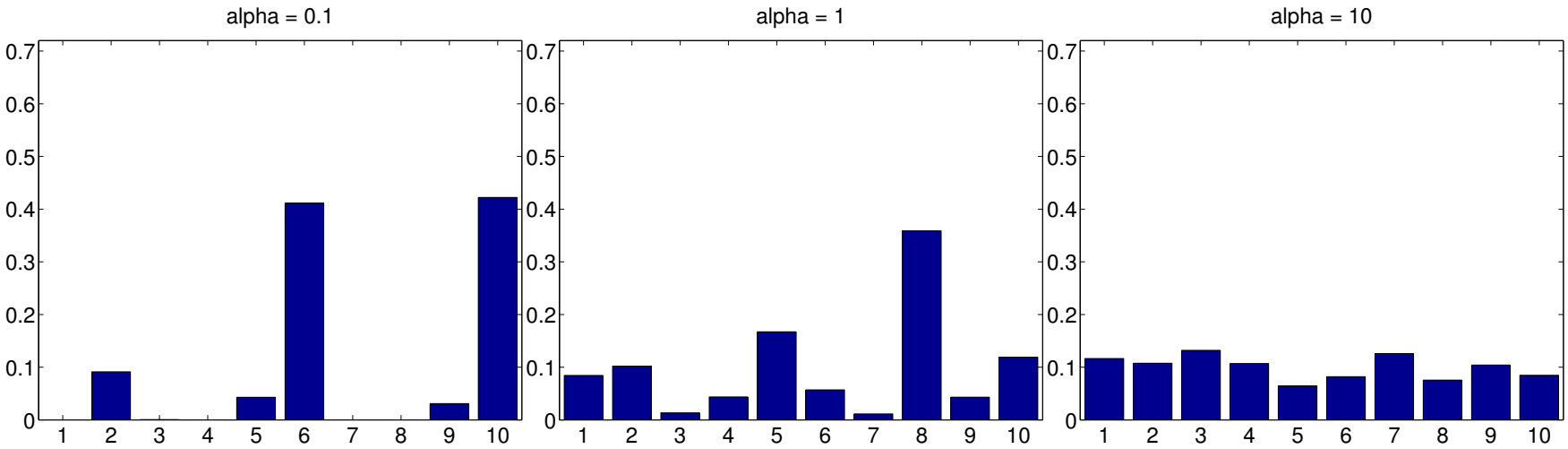


The symmetric Dirichlet distribution

In the symmetric Dirichlet distribution all parameters are identical: $\alpha_i = \alpha, \quad \forall i$.

`en.wikipedia.org/wiki/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif`

To sample from a symmetric Dirichlet in D dimensions with concentration α use:
`w = randg(alpha,D,1); bar(w/sum(w));`



Distributions drawn at random from symmetric 10 dimensional Dirichlet distributions with various concentration parameters.