Bayesian ML : Posterior Distributions and Mixture Models

Suresh Manandhar

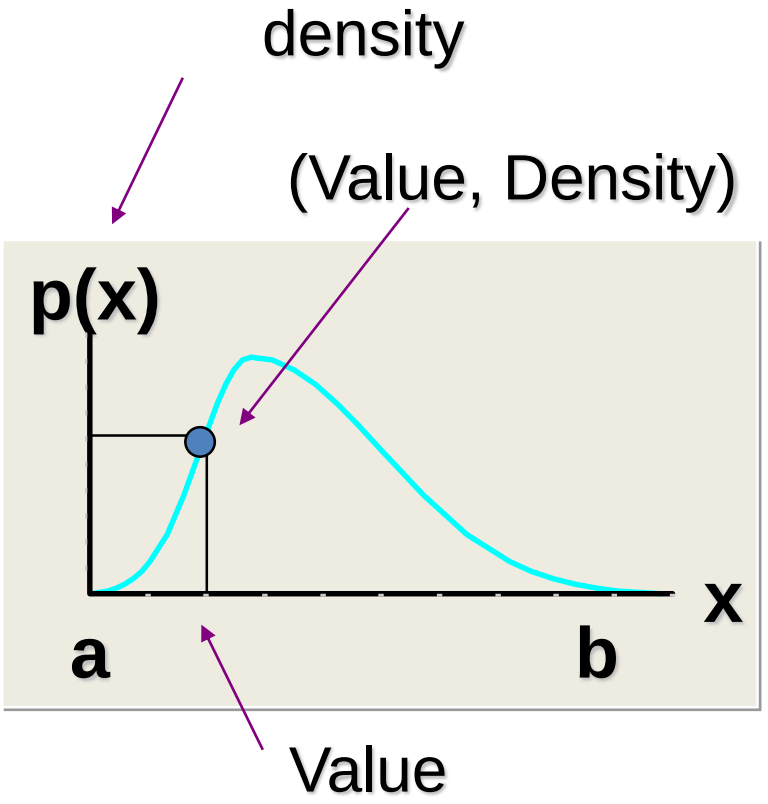
Naamii, Nepal

Continuous Probability Density Function (pdf)

- Shows all values of x in the given interval $[a, b]$, the **density** $p(x)$
- $p(x)$ is a ***probability density function*** (**pdf**)

- Since probabilities need to sum to 1, the corresponding condition is:

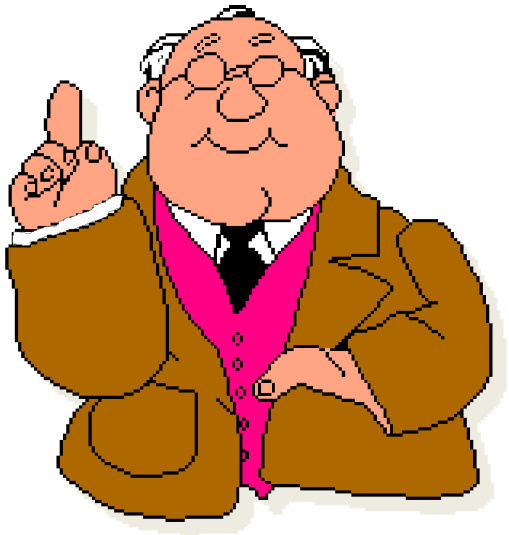
$$F(-\infty \leq X \leq \infty)$$
$$= \int_{-\infty}^{\infty} p(x)dx = 1$$



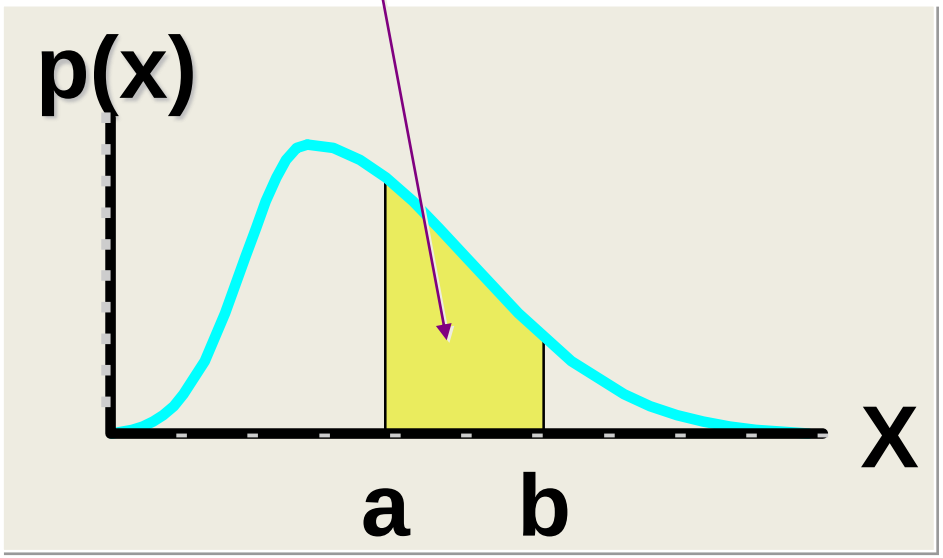
Cumulative density (cdf)

The probability that x lies in the interval $[a, b]$ is given by $F(a \leq X \leq b)$.

Cumulative probability is Area Under Curve!



$$F(a \leq X \leq b) = \int_a^b p(x)dx$$



Some properties

$$F(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} p(x)dx = 1$$

$$F(X = a) = F(a \leq X \leq a) = \int_{a}^{a} p(x)dx = 0$$

$$F(a) = F(-\infty \leq X \leq a) = \int_{-\infty}^{a} p(x)dx$$

- or, more generally:

$$F(x) = \int p(x)dx$$

- $F(x)$ is known as the ***cumulative distribution function***. (**CDF**).
- The pdf and cdf are related by:

$$\frac{d}{dx}F(x) = p(x)$$

- The following provides an intuitive understanding:

$$F\left(a - \frac{\Delta}{2} \leq X \leq a + \frac{\Delta}{2}\right) = \int_{a-\frac{\Delta}{2}}^{a+\frac{\Delta}{2}} p(x)dx \simeq \Delta\, \mathrm{p(a)}$$

Expectation and Variance

Weighted Average

$$E[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

Averaged Squared Distance From Mean

$$Var(X) = \int_{-\infty}^{\infty} (x - E[X])^2\, p(x)\, dx$$

Bayes for Continuous random variables

Law of Total Probability

- For **discrete** random variables:

$$p(X = x) = \sum_i p(X = x, Y = y_i)$$

$$= \sum_i p(X = x | Y = y_i) p(Y = y_i)$$

- For **continuous** random variables:

$$p(x) = \int p(x, y) dy$$

$$= \int p(x|y)\, p(y) dy$$

Bayes for Continuous random variables

- From the definition of conditional probability:

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)}$$

- Substituting for $p(x)$ we derive.

Bayes' theorem for continuous random variables:

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)\, p(y)dy}$$

Some continuous probability distributions

Distribution	Support/Range
Uniform distribution	$(\boldsymbol{a}, \boldsymbol{b})$
Gaussian (or Normal) distribution	$(-\infty, +\infty)$
Exponential distribution	$(\mathbf{0}, +\infty]$
Beta distribution	$[\mathbf{0}, \mathbf{1}]$
Dirichlet distribution	$[\mathbf{0}, \mathbf{1}]^{\boldsymbol{n}}$
Multivariate Gaussian distribution	$(-\infty, +\infty)^{\boldsymbol{n}}$

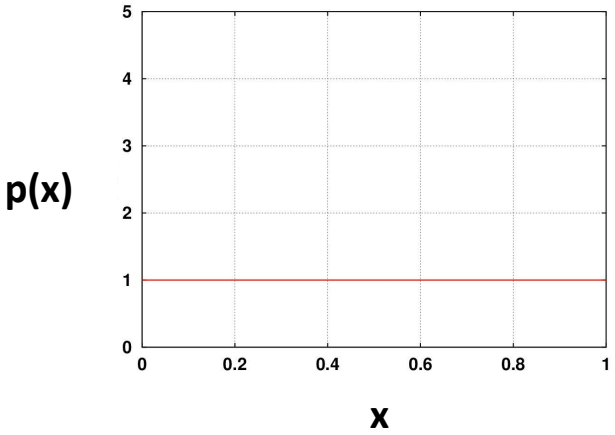
Uniform Distribution

- **Definition**: A random variable is uniformly distributed in the interval $(\mathbf{0}, \mathbf{1})$ if:

$$\boldsymbol{p(x)} = \begin{cases} \mathbf{1} & if \quad 0 < x < 1 \\ \mathbf{0} & otherwise \end{cases}$$

- Hence:

$$\int_{-\infty}^{\infty} \boldsymbol{p(x)dx} = \int_{\mathbf{0}}^{\mathbf{1}} \mathbf{1}\, \boldsymbol{dx} = \boldsymbol{x}\Big|_{\mathbf{0}}^{\mathbf{1}} = \mathbf{1} - \mathbf{0} = \mathbf{1}$$



Uniform Distribution

- **Definition**: A random variable is uniformly distributed in the interval $(\mathbf{0}, \mathbf{1})$ if:

$$\boldsymbol{p(x)} = \begin{cases} \mathbf{1} & if \quad 0 < x < 1 \\ \mathbf{0} & otherwise \end{cases}$$

- Hence:

$$\int_{-\infty}^{\infty} \boldsymbol{p(x)dx} = \int_{\mathbf{0}}^{\mathbf{1}} \mathbf{1}\, \boldsymbol{dx} = \boldsymbol{x}\Big|_{\mathbf{0}}^{\mathbf{1}} = \mathbf{1} - \mathbf{0} = \mathbf{1}$$

- Then for $\boldsymbol{a} \leq \boldsymbol{X} \leq \boldsymbol{b}$:

$$\boldsymbol{F(0 < a \leq X \leq b < 1)} = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \mathbf{1}\, \boldsymbol{dx} = \boldsymbol{x}\Big|_{\boldsymbol{a}}^{\boldsymbol{b}} = \boldsymbol{b} - \boldsymbol{a}$$

Uniform Distribution

- **Definition**: A random variable is uniformly distributed in the interval $(\mathbf{0}, \mathbf{1})$ if:

$$\boldsymbol{p(x)} = \begin{cases} \mathbf{1} & if \quad 0 < x < 1 \\ \mathbf{0} & otherwise \end{cases}$$

- Hence:

$$\int_{-\infty}^{\infty} \boldsymbol{p(x)dx} = \int_{\mathbf{0}}^{\mathbf{1}} \mathbf{1}\, \boldsymbol{dx} = \boldsymbol{x}\Big|_{\mathbf{0}}^{\mathbf{1}} = \mathbf{1} - \mathbf{0} = \mathbf{1}$$

- **Definition**: A random variable is uniformly distributed in the interval $(\boldsymbol{a}, \boldsymbol{b})$ if:

$$\boldsymbol{p(x)} = \begin{cases} \dfrac{\mathbf{1}}{\boldsymbol{b} - \boldsymbol{a}} & if \quad a < x < b \\ \mathbf{0} & otherwise \end{cases}$$

- Hence:

$$\int_{-\infty}^{\infty} \boldsymbol{p(x)dx} = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \frac{\mathbf{1}}{\boldsymbol{b} - \boldsymbol{a}}\, \boldsymbol{dx} = \frac{\boldsymbol{x}}{\boldsymbol{b} - \boldsymbol{a}}\Big|_{\boldsymbol{a}}^{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\boldsymbol{b} - \boldsymbol{a}} - \frac{\boldsymbol{a}}{\boldsymbol{b} - \boldsymbol{a}} = \mathbf{1}$$

Example

- $\boldsymbol{X}$ is uniformly distributed over $(\mathbf{0}, \mathbf{10})$ calculate:

$$\boldsymbol{F(X < 3)}, \quad \boldsymbol{F(X > 3)}, \quad \boldsymbol{F(1 < X < 6)}$$

Example

- $\boldsymbol{X}$ is uniformly distributed over $(\mathbf{0}, \mathbf{10})$ calculate:

$$\boldsymbol{F(X < 3)}, \quad \boldsymbol{F(X > 3)}, \quad \boldsymbol{F(1 < X < 6)}$$

- **Solution**:

$$\boldsymbol{p(x)} = \begin{cases} \dfrac{\mathbf{1}}{\mathbf{10}} & if \quad 0 < x < 10 \\ \mathbf{0} & otherwise \end{cases}$$

$$\boldsymbol{F(X < 3)} = \int_{-\infty}^{\mathbf{3}} \boldsymbol{p(x)\, dx} = \frac{\int_{\mathbf{0}}^{\mathbf{3}} \boldsymbol{dx}}{\mathbf{10}} = \frac{\mathbf{3}}{\mathbf{10}}$$

$$\boldsymbol{F(X > 3)} = \int_{\mathbf{3}}^{+\infty} \boldsymbol{p(x)\, dx} = \frac{\int_{\mathbf{3}}^{\mathbf{10}} \boldsymbol{dx}}{\mathbf{10}} = \frac{\mathbf{7}}{\mathbf{10}}$$

$$\boldsymbol{F(1 < X < 6)} = \int_{\mathbf{1}}^{\mathbf{6}} \boldsymbol{p(x)\, dx} = \frac{\int_{\mathbf{1}}^{\mathbf{6}} \boldsymbol{dx}}{\mathbf{10}} = \frac{\mathbf{5}}{\mathbf{10}} = \frac{\mathbf{1}}{\mathbf{2}}$$

Normal Distribution

- The normal (or Gaussian) distribution gives the familiar bell shaped curve
- **Definition**: A random variable $\boldsymbol{X}$ is normally distributed if its probability density function is given by:

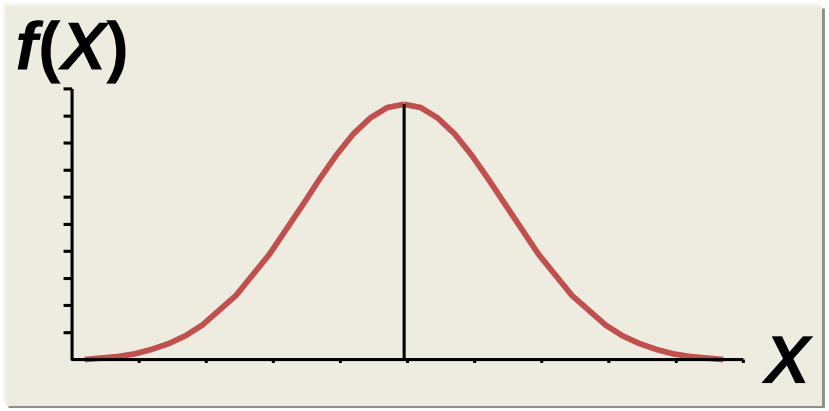
$$\boldsymbol{p(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

- The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ are known as the mean and variance.
- It turns out that:

$$\boldsymbol{E[X] = \mu}$$
$$\boldsymbol{\mathbf{Var}(X) = \sigma^2}$$

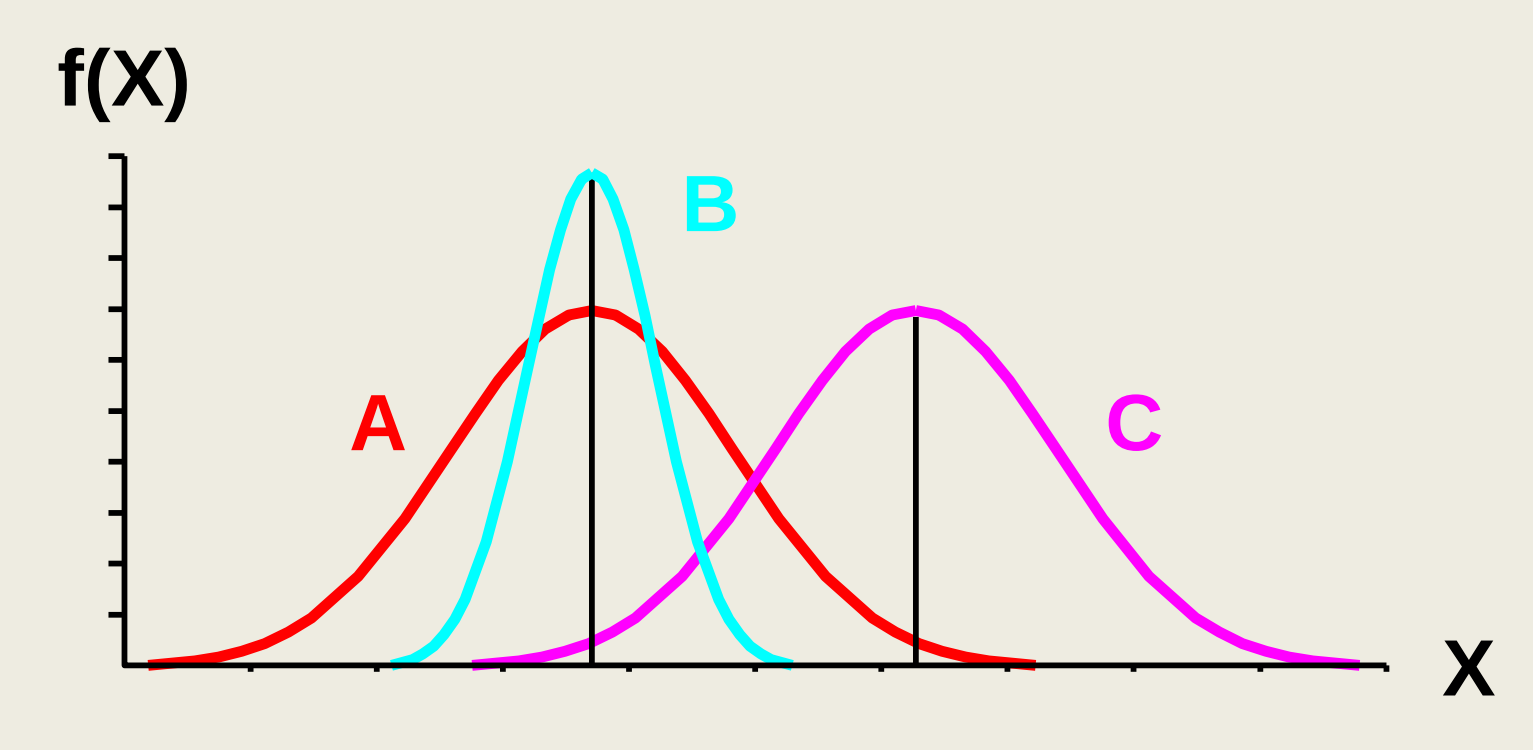
The Normal Distribution

- 1. ‘Bell-Shaped’ & Symmetrical
- 2. Mean, median, mode are equal
- 3. Random variable has infinite range



Mean
Median
Mode

Effect of varying μ and σ^2



Standardize the Normal Distribution

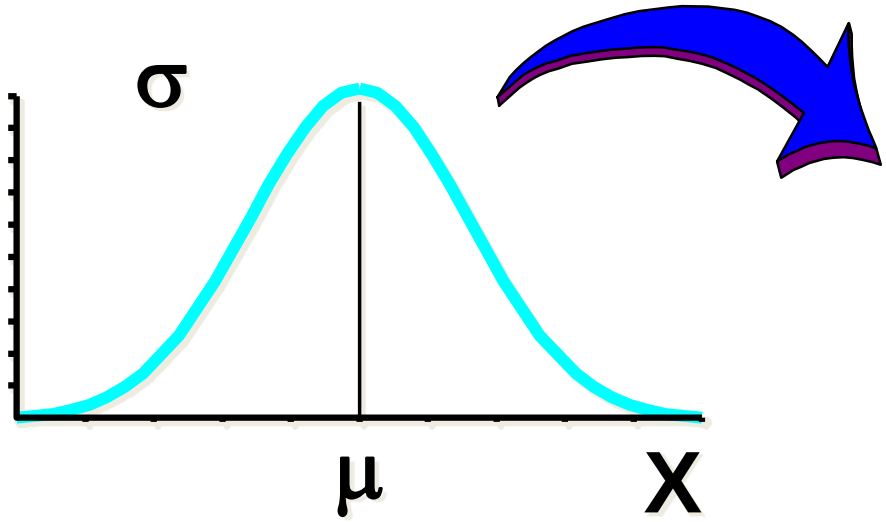
$$p(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

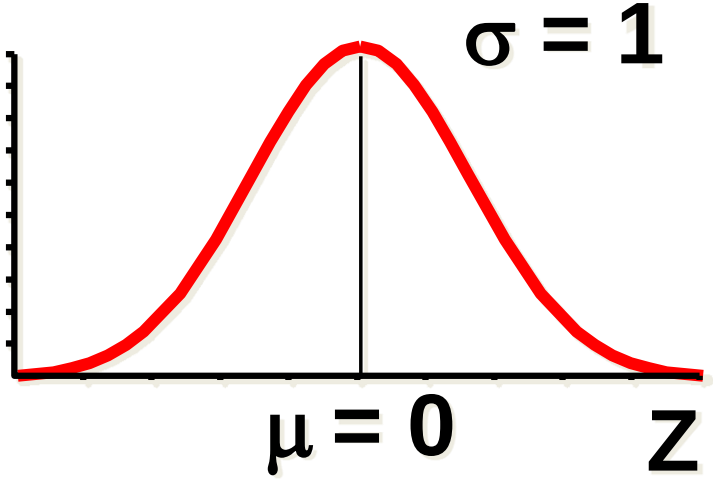
$$p(X = x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Normal Distribution

$$Z = \frac{X - \mu}{\sigma}$$

Standardized Normal Distribution





One table!

Some properties of Normal distribution

- **Central Limit Theorem**
  - *The sum of a set of independent random variables approaches the* ***Gaussian distribution*** *as the number of variables* $\rightarrow\infty$*,* regardless of the distributions of the individual variables
  - (There are generalisations of the **CLT**.)
  - **Example**: the sum of the face value of randomly drawn playing cards has an approximately Gaussian distribution
- Simplicity: specified by only two intuitive parameters
- Mathematically tractable
  - Many analyses turn out very simply with the Normal distribution

Beta Distribution

- Suppose that the coin factory makes coins that are not perfectly fair all the time
- So, most times the coins are fair but less often the coins are slightly unfair, and less-less often the coins are quite unfair
- So, there is a distribution over the bias of the coins
- **Question**: How to model this distribution?
- The output/sample from this distribution will be the **bias** of a coin i.e. a number between $\mathbf{0}$ and $\mathbf{1}$

Gamma function

- The gamma function $\boldsymbol{\Gamma(N)}$ generalises the factorial function to the reals such that:

  $\boldsymbol{\Gamma(N+1) = N!}$ for natural number $\boldsymbol{N}$

  $\boldsymbol{\Gamma(1) = 0! = 1}$

  $\boldsymbol{\Gamma\left(\frac{1}{2}\right) = \left(-\frac{1}{2}\right)! = \sqrt{\pi}}$

  $\boldsymbol{\Gamma(0) = (-1)! = \frac{\pi}{2}}$

- The gamma function can be used to ***interpolate*** for values for which the factorial is undefined.

Beta distribution

$$\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha-1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta-1}}}{\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})} \sim \boldsymbol{Beta}(\boldsymbol{\alpha},\boldsymbol{\beta})$$

- The $-\mathbf{1}$'s can be thought as mathematical convenience/convention
- To be able to show that such a thing exists, we need to show that:

$$\int_{\mathbf{0}}^{\mathbf{1}} \mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta})\boldsymbol{d\theta} = \int_{\mathbf{0}}^{\mathbf{1}} \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha-1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta-1}}}{\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})}\boldsymbol{d\theta} = \mathbf{1}$$

- Equivalently, need to show that $\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is well defined:

$$\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \int_{\mathbf{0}}^{\mathbf{1}} \boldsymbol{\theta}^{\boldsymbol{\alpha-1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta-1}}\boldsymbol{d\theta}$$

Beta function - derivation

Beta function: $B(\alpha,\beta)=\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$

- We will apply '*integration by parts*', with $u=\theta^{\alpha-1}$, and, $dv=(1-\theta)^{\beta-1}d\theta$ so the integral becomes

$$B(\alpha,\beta)=\int_0^1 udv=uv-\int_0^1 vdu$$

- $du=(\alpha-1)\theta^{\alpha-2}d\theta$, and, $v=-\frac{1}{\beta}(1-\theta)^{\beta}$. Thus:

$$=\theta^{\alpha-1}\left(-\frac{1}{\beta}(1-\theta)^{\beta}\right)\Big|_0^1-\int_0^1\left(-\frac{1}{\beta}(1-\theta)^{\beta}\right)(\alpha-1)\theta^{\alpha-2}d\theta$$

$$=\frac{(\alpha-1)}{\beta}\int_0^1(1-\theta)^{\beta}\theta^{\alpha-2}d\theta=\frac{(\alpha-1)}{\beta}B(\alpha-1,\beta+1)$$

$$=\frac{(\alpha-1)(\alpha-2)\dots 1}{\beta(\beta+1)\dots(\beta+\alpha-2)}B(1,\beta+\alpha-1)$$

$$=\frac{(\alpha-1)(\alpha-2)\dots 1}{\beta(\beta+1)\dots(\beta+\alpha-2)}\int_0^1\theta^{1-1}(1-\theta)^{\beta+\alpha-2}d\theta$$

$$= \frac{(\alpha-1)(\alpha-2)\dots 1}{\beta(\beta+1)\dots(\beta+\alpha-2)} \int_0^1 (1-\theta)^{\beta+\alpha-2} d\theta$$

$$= \frac{(\alpha-1)(\alpha-2)\dots 1}{\beta(\beta+1)\dots(\beta+\alpha-2)} \left(-\frac{(1-\theta)^{\beta+\alpha-1}}{\beta+\alpha-1} \right) \Bigg|_0^1$$

$$= \frac{(\alpha-1)(\alpha-2)\dots 1}{\beta(\beta+1)\dots(\beta+\alpha-2)(\beta+\alpha-1)}$$

$$= \frac{\Gamma(\alpha)}{\beta(\beta+1)\dots(\beta+\alpha-2)(\beta+\alpha-1)}$$

$$= \frac{1\dots(\beta-2)(\beta-1)\Gamma(\alpha)}{1\dots(\beta-2)(\beta-1)\beta(\beta+1)\dots(\beta+\alpha-2)(\beta+\alpha-1)}$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

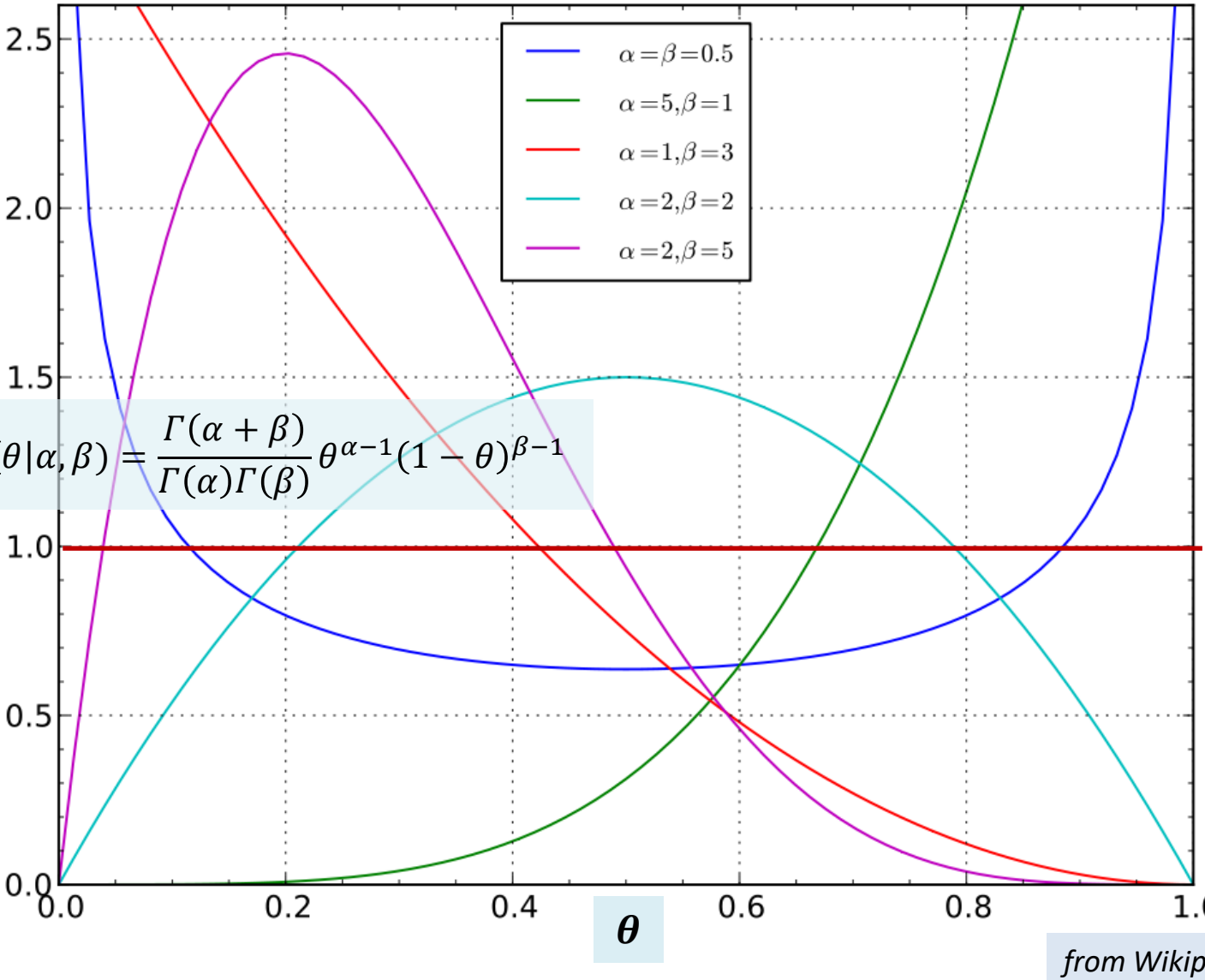
Strictly speaking this proof only applies to positive natural values of α, β

- **Thus,** $B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ **hence:**

$$\mathrm{p}(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- For $\alpha = \beta = 1$, the distribution is ***uniform*** with $\mathrm{p}(\theta|\alpha,\beta) = 1$

Shape of Beta distribution



$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

from Wikipedia

Modelling the coin factory

- If the factory has a high quality control then choosing $\boldsymbol{\alpha = \beta = 100}$ will give a very peaky pdf
- What would choosing a uniform distribution, $\boldsymbol{\alpha = \beta = 1}$, give? Would that be a good factory?
- Similarly, what about values less than 1?

Modelling typical coins from coin factory

- Suppose, the factory inspector visits the coin factory and picks a coin at random from the factory
- Remember, the factory has beta parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$
- The inspector flips the coin $\boldsymbol{N}$ times and sees

$$\boldsymbol{c} = (\boldsymbol{c_1}, \boldsymbol{c_2})$$

i.e $\boldsymbol{c_1}$ heads and $\boldsymbol{c_2}$ tails with $\boldsymbol{N} = \boldsymbol{c_1} + \boldsymbol{c_2}$

- What is the probability of $\boldsymbol{c_1}$ heads and $\boldsymbol{c_2}$ tails?
- How do we compute this?
- Can we use Bayes?

N here equals $\boldsymbol{c_1} + \boldsymbol{c_2}$

Modelling typical coins from coin factory

- What we want is: $\boldsymbol{p(c|\alpha,\beta)}$**?**

- How do we derive this: (*use law of total probability*)

$$\boldsymbol{p(c|\alpha,\beta)} = \int \boldsymbol{p(c,\theta|\alpha,\beta)}\, \boldsymbol{d\theta}$$

$$= \int \boldsymbol{p(c|\theta)}\, \boldsymbol{p(\theta|\alpha,\beta)}\, \boldsymbol{d\theta}$$

Binomial Likelihood ***Beta Prior***

- Integration can be solved analytically (i.e. by hand) if the two distributions have similar form.
- In this case, we say that that two distributions are **conjugate**.

Modelling coin tosses

- Once, we are happy with the choice of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ for our factory we can plug this in

$$p(c|\alpha,\beta) = \int p(c,\theta|\alpha,\beta)\, d\theta = \int p(c|\theta)\, p(\theta|\alpha,\beta)\, d\theta$$

$$= \int \frac{N!}{c_1!\, c_2!}\, \theta^{c_1}(1-\theta)^{c_2}\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

$$= \frac{N!}{c_1!\, c_2!}\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{c_1}(1-\theta)^{c_2}\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

$$= \frac{N!}{c_1!\, c_2!}\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1} d\theta$$

conjugacy helps

$$= \frac{N!}{c_1!\, c_2!}\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1} d\theta$$

$$= \frac{N!}{c_1!\, c_2!}\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \frac{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}{\Gamma(c_1+c_2+\alpha+\beta)}$$

from definition of Beta function

N here equals $c_1 + c_2$

The Beta-Binomial Distribution

- What we have just derived is the beta-binomial distribution that gives the (averaged) probability of drawing c_1 heads and c_2 tails from a coin that has been drawn from a beta distribution with parameters α and β.
- The 'averaged' above means that the coin parameter θ has been ***integrated out*** (and hence no longer appears in the equations):

$$p(c|\alpha,\beta) = \frac{N!}{c_1!\,c_2!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}{\Gamma(c_1+c_2+\alpha+\beta)}$$

- Using the definition of the ***beta function***:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$p(c|\alpha,\beta) = \frac{N!}{c_1!\,c_2!}\frac{B(c_1+\alpha,c_2+\beta)}{B(\alpha,\beta)}$$

- $\frac{N!}{c_1!c_2!} \neq \frac{1}{B(C_1+1,C_2+1)}$ since $\frac{1}{B(C_1+1,C_2+1)} = \frac{\Gamma(C_1+C_2+2)}{\Gamma(C_1+1)\Gamma(C_2+1)} = \frac{(N+1)!}{C_1!C_2!}$

Inference

Inference

- The inference problem can be viewed as determining the parameters of your model from observations
- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair?

Inference

- The inference problem can be viewed as determining the parameters of your model from observations
- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair? Or what is the probability that it is fair? How do you even ask such a question?

Inference

- The inference problem can be viewed as determining the parameters of your model from observations
- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair? Or what is the probability that it is fair? How do you even ask such a question?
- The key to Bayesian inference is the mechanism to integrate our prior beliefs into the modelling process and provide mathematically grounded answers to the above questions.

Inference

- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair?

$$\boldsymbol{p(c = (c_1 = 10, c_2 = 10)|\theta) = \binom{20}{10} \theta^{10}(1-\theta)^{10}}$$

- What is $\boldsymbol{p(\theta|c = (c_1 = 10, c_2 = 10))}$?
- Using the definition of conditional probability:

$$\boldsymbol{p(\theta, c) = p(\theta|c)\, p(c) = p(c|\theta)p(\theta)}$$

- Using Bayes':

$$\boldsymbol{p(c) = \int_0^1 p(\theta, c)d\theta = \int_0^1 p(c|\theta)p(\theta)d\theta}$$

- Substituting, we get:

$$\boldsymbol{p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}}$$

Inference

- Thus to find out $\boldsymbol{p(\theta|c = (c_1 = 10, c_2 = 10))}$ we can use:

$$\boldsymbol{p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}}$$

- However, we do not know what $\boldsymbol{p(\theta)}$ is?
- Hmmm… What might this mean?

Inference

- Thus to find out $\boldsymbol{p(\theta|c = (c_1 = 10, c_2 = 10))}$ we can use:

$$\boldsymbol{p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}}$$

- However, we do not know what $\boldsymbol{p(\theta)}$ is?
- Hmmm… What might this mean?
- $\boldsymbol{p(\theta)}$ is our prior belief about the coin with bias $\boldsymbol{\theta}$
- What does this mean?

Inference

$$p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int p(c|\theta)p(\theta)d\theta}$$

Posterior:

- $p(\theta|c)$ is the ***posterior distribution***
- Gives the probability of the parameter given data

Prior:

- $p(\theta)$ is the ***prior distribution***
- Gives the probability for different values of θ and quantifies our belief regarding θ

Likelihood:

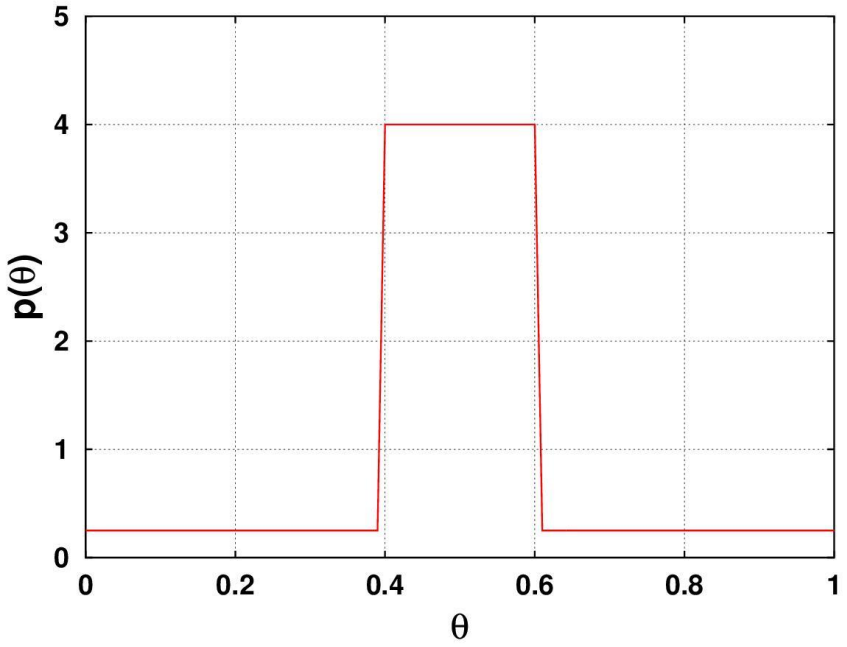
- $p(c|\theta)$ is the ***likelihood*** of the data given by the model
- Gives the probability of the data being generated by the model

Partition function/Normalising constant/Evidence:

- $\int_0^1 p(c|\theta)p(\theta)d\theta$ is the partition function/normalising constant/evidence.
- This is a constant needed to ensure that the probabilities sum to 1.

Inference

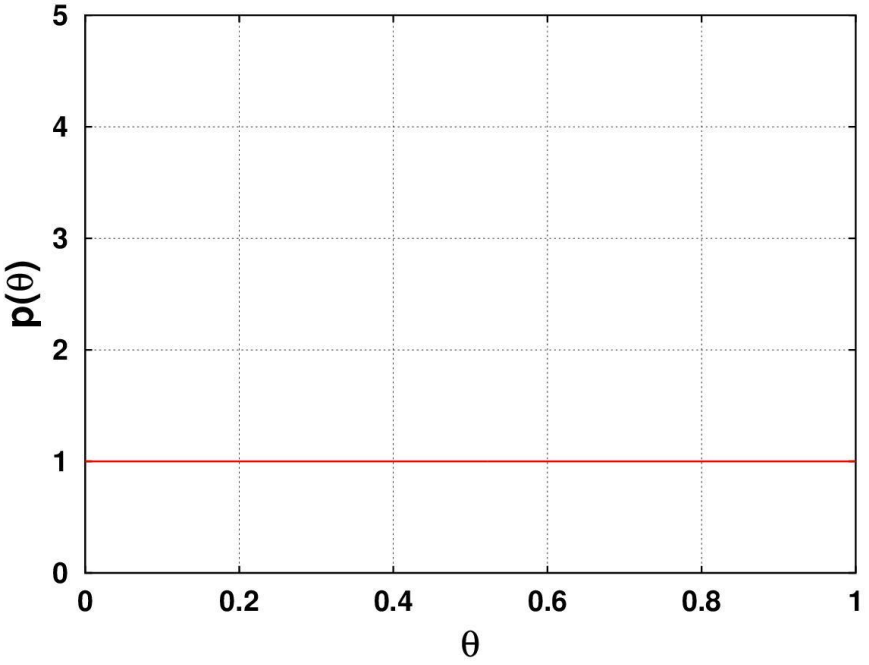
- $\boldsymbol{p(\theta)}$ is our prior belief about the coin with bias $\boldsymbol{\theta}$
- Suppose, we say that, $\boldsymbol{p(\theta)}$ is **piecewise uniform** with:
  - $\boldsymbol{F(0.4 \leq \theta \leq 0.6) = 0.8}$, $\boldsymbol{F(0 \leq \theta < 0.4) = 0.1}$, $\boldsymbol{F(0.6 < \theta \leq 1) = 0.1}$



Inference

- Alternatively suppose that, $\boldsymbol{p(\theta)}$ is **uniform** with:

$$\boldsymbol{p(\theta)} = \begin{cases} \mathbf{1} & if \quad 0 < \theta < 1 \\ \mathbf{0} & otherwise \end{cases}$$



Inference

- Alternatively still, we can assume that $\boldsymbol{p(\theta)}$ is given by a ***beta distribution*** with parameters $\boldsymbol{\alpha, \beta}$

$$\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha}-\mathbf{1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta}-\mathbf{1}}}{\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})} = \frac{\boldsymbol{\Gamma}(\boldsymbol{\alpha}+\boldsymbol{\beta})}{\boldsymbol{\Gamma}(\boldsymbol{\alpha})\boldsymbol{\Gamma}(\boldsymbol{\beta})}\boldsymbol{\theta}^{\boldsymbol{\alpha}-\mathbf{1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta}-\mathbf{1}}$$

- The uniform distribution can be recovered by $\boldsymbol{\alpha = 1, \beta = 1}$

$$\mathbf{p}(\boldsymbol{\theta}|\mathbf{1},\mathbf{1}) = \frac{\boldsymbol{\theta}^{\mathbf{1}-\mathbf{1}}(\mathbf{1}-\boldsymbol{\theta})^{\mathbf{1}-\mathbf{1}}}{\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})} = \frac{\boldsymbol{\Gamma}(\mathbf{1}+\mathbf{1})}{\boldsymbol{\Gamma}(\mathbf{1})\boldsymbol{\Gamma}(\mathbf{1})} = \mathbf{1}$$

- The piecewise uniform distribution may also be approximated by suitable choices of $\boldsymbol{\alpha, \beta}$
- Choosing the ***beta distribution*** as a **prior** means that we can take advantage of ***conjugacy***.

Inference: Binomial Posterior Distribution

- For the case when $p(\theta)$ is given by the **beta distribution** with parameters α, β:

$$p(\theta|c,\alpha,\beta) = \frac{p(c|\theta,\alpha,\beta)\, p(\theta|\alpha,\beta)}{\int_0^1 p(c|\theta,\alpha,\beta)\, p(\theta|\alpha,\beta) d\theta}$$

Remember $c = (c_1, c_2)$

$$= \frac{\frac{N!}{c_1!\, c_2!}\, \theta^{c_1}(1-\theta)^{c_2}\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \frac{N!}{c_1!\, c_2!}\, \theta^{c_1}(1-\theta)^{c_2}\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}$$

$$= \frac{\theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1}}{\frac{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}{\Gamma(c_1+c_2+\alpha+\beta)}}$$

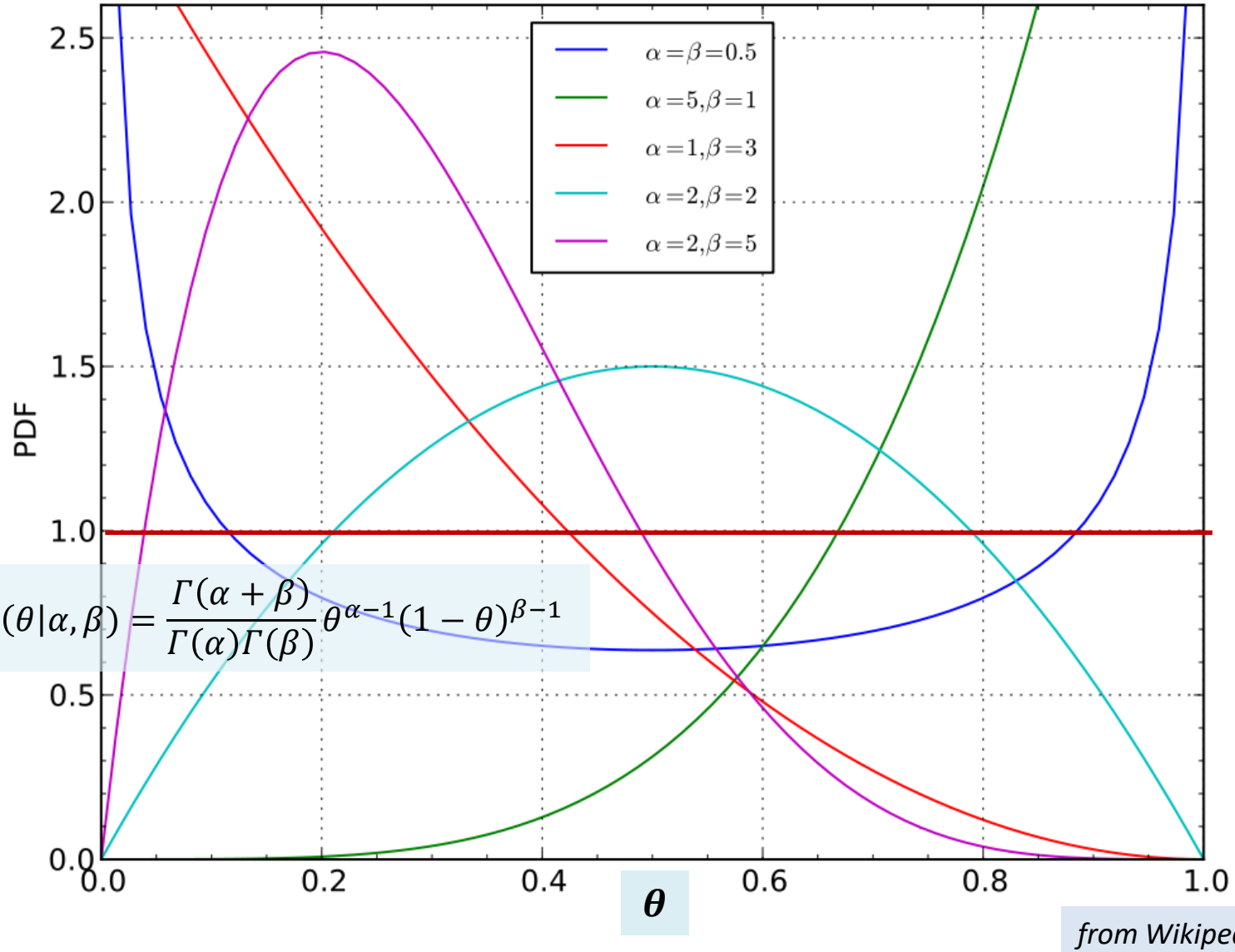
$$= \frac{\Gamma(c_1+c_2+\alpha+\beta)}{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}\, \theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1}$$

$$= \frac{1}{B(c_1+\alpha, c_2+\beta)}\, \theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1}$$

$$= p(\theta|\alpha+c_1, \beta+c_2) \sim Beta(\alpha+c_1, \beta+c_2)$$

- Thus, posterior of binomial (under beta) is beta with counts added into the parameters

Shape of Beta distribution



$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

from Wikipedia

Aside: Inference

- If $\boldsymbol{\theta}$ is ***piecewise uniform*** with, $\boldsymbol{F(0.4 \leq \theta \leq 0.6) = 0.8}$, $\boldsymbol{F(0 \leq \theta < 0.4) = 0.1}$, $\boldsymbol{F(0.6 < \theta \leq 1) = 0.1}$
- We can calculate, $\boldsymbol{p(\theta)}$:

$\boldsymbol{F(0.4 \leq \theta \leq 0.6) = 0.8}$ implies $\boldsymbol{p(\theta) = \frac{0.8}{0.2} = 4}$

$\boldsymbol{F(0 \leq \theta < 0.4) = 0.1}$ implies $\boldsymbol{p(\theta) = \frac{0.1}{0.4} = \frac{1}{4}}$

$\boldsymbol{F(0.6 < \theta \leq 1) = 0.1}$ implies $\boldsymbol{p(\theta) = \frac{0.1}{0.4} = \frac{1}{4}}$

Aside: Inference

$$p(\theta|c) = \frac{p(c|\theta)\,p(\theta)}{\int_0^1 p(c|\theta)\,p(\theta)d\theta}$$

$$= \frac{p(c|\theta)\,p(\theta)}{\int_0^{0.4} p(c|\theta)\,.\frac{1}{4}\,d\theta + \int_{0.4}^{0.6} p(c|\theta).4\,d\theta + \int_{0.6}^{1} p(c|\theta)\,.\frac{1}{4}\,d\theta}$$

$$= \frac{\theta^{c_1}(1-\theta)^{c_2}\,p(\theta)}{\frac{1}{4}IB(.4,c_1,c_2) + 4[IB(.6,c_1,c_2) - IB(.4,c_1,c_2)] + \frac{1}{4}[B(c_1,c_2) - IB(.6,c_1,c_2)]}$$

- Where IB is the ***incomplete beta function*** defined by:

$$IB(a,c_1,c_2) = \int_0^a \theta^{c_1-1}(1-\theta)^{c_2-1}d\theta \qquad 0 \le a \le 1$$

- Hence: $IB(1,c_1,c_2) = B(c_1,c_2)$
- Most math packages provide implementations of the incomplete beta function.
- Thus, posterior probability for any θ and c can be calculated.

Bayesian Inference

$$p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int p(c|\theta)p(\theta)d\theta}$$

- ***Being Bayesian*** typically means that we treat $p(\theta|c)$ as a distribution
- This means that we do not get a single value for the model parameter θ
- However, sometimes, we may want to find out what the '**best**' value for the model parameter θ would be.
- How do define what '**best**' means ?

MLE Inference

- This is the maximum likelihood estimate (where c is the data)

$$\boldsymbol{\theta_{MLE} = arg\,\max_{\theta} p(c|\theta)}$$

- Thus MLE is simply the maximum/**mode** of the model likelihood.
- We don't even need to define a prior, so (apparently) more "objective".
- Typically deal with situations where there is only one mode, so can talk about **the** MLE.

MAP Inference

$$p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int p(c|\theta)p(\theta)d\theta}$$

MAP (Maximum a Posteriori):

- Assume that $p(\theta)$ is distributed as per ***some given distribution***

$$\theta_{MAP} = arg\,max_{\theta}\, p(\theta|c) = arg\,max_{\theta} \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}$$

- Since the denominator is a constant:

$$\theta_{MAP} = arg\,max_{\theta}\, p(c|\theta)p(\theta)$$

- Thus MAP estimate is the ***mode*** of the posterior distribution

Multinomial Distribution

- Probability of observing $\boldsymbol{c} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_k})$ heads in all possible ways out of $\boldsymbol{C} = \sum_i \boldsymbol{c_i}$ throws from a k-headed dice with probability of heads $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_k})$ *s.t.* $\sum_i \boldsymbol{\theta_i} = \mathbf{1}$

$$\boldsymbol{p(heads = c|\theta)} = \frac{\boldsymbol{C}!}{\prod_i \boldsymbol{c_i}!} \prod_{\boldsymbol{i}} \boldsymbol{\theta^{c_i}}$$

- Points to note:
  - Generalises the binomial distribution to $\boldsymbol{k} > \mathbf{2}$
  - Equivalent to the binomial for $\boldsymbol{k} = \mathbf{2}$

The Dirichlet Distribution

- Dirichlet distribution generalises the Beta distribution to the $\boldsymbol{k-1}$ probability simplex
- The Beta distribution:

$$\boldsymbol{p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \sim Beta(\alpha,\beta)}$$

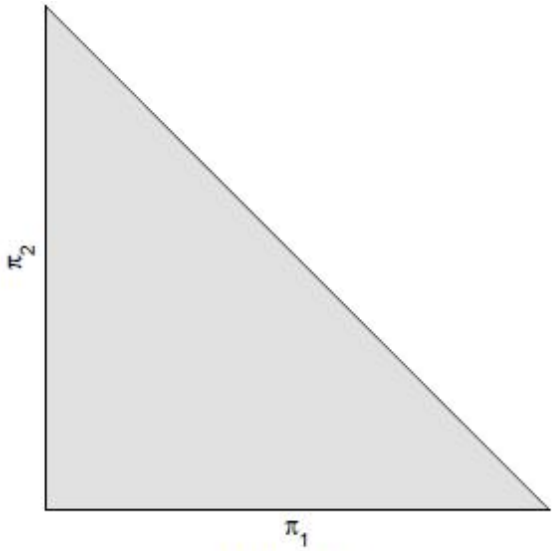
- The Dirichlet distribution:

$$\boldsymbol{p(\theta|\alpha) = \frac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i-1} \sim Dir(\alpha_1,\ldots,\alpha_k)}$$

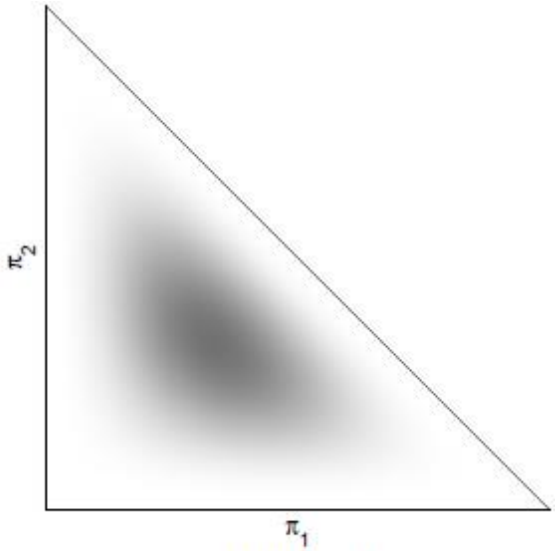
with:

$$\boldsymbol{\alpha = (\alpha_1,\ldots,\alpha_k),\ \theta = (\theta_1,\ldots,\theta_k),\ \sum_i \theta_i = 1,\ A = \sum_i \alpha_i}$$

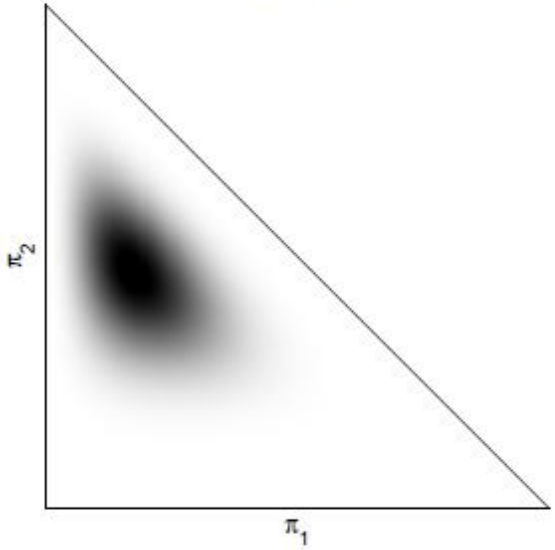
- So, the samples from the Dirichlet distribution can be used to model the bias in a $\boldsymbol{k}$-sided dice.



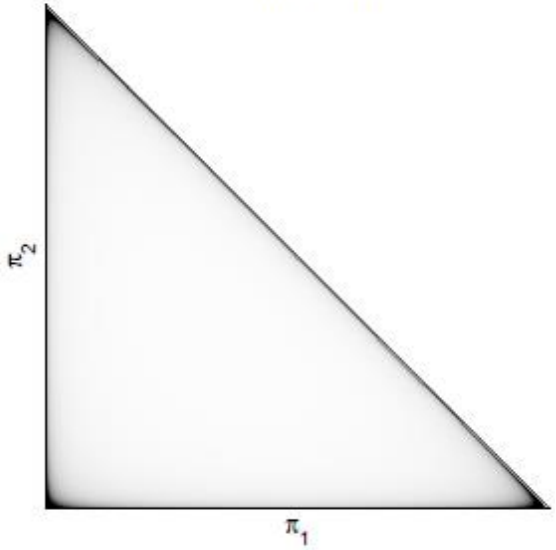
$\pi \sim \mathrm{Dir}(1, 1, 1)$



$\pi \sim \mathrm{Dir}(4, 4, 4)$



$\pi \sim \mathrm{Dir}(4, 9, 7)$



$\pi \sim \mathrm{Dir}(0.2, 0.2, 0.2)$

The Dirichlet-Multinomial

- Like for the Beta-Binomial distribution we can integrate out the Multinomial parameters
- Here, we are modelling the case where we need to predict the outcome $\boldsymbol{c} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_k})$ of a dice throw from a dice sampled from a factory with parameter $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_k})$

$$\boldsymbol{p(c|\alpha)} = \int \boldsymbol{p(c, \theta|\alpha)}\, \boldsymbol{d\theta} = \int \boldsymbol{p(c|\theta)}\, \boldsymbol{p(\theta|\alpha)}\, \boldsymbol{d\theta}$$

$$= \int \left[\frac{\boldsymbol{C!}}{\prod_{\boldsymbol{i}} \boldsymbol{c_i!}} \prod_{\boldsymbol{i}} \boldsymbol{\theta}^{\boldsymbol{c_i}} \right] \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{\alpha_i - 1}}\, \boldsymbol{d\theta}$$

$$= \frac{\boldsymbol{C!}}{\prod_{\boldsymbol{i}} \boldsymbol{c_i!}} \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \int \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{c_i + \alpha_i - 1}}\, \boldsymbol{d\theta}$$

- From earlier: $\boldsymbol{B(\alpha, \beta)} = \int_{\boldsymbol{0}}^{\boldsymbol{1}} \boldsymbol{\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta} = \frac{\boldsymbol{\Gamma(\alpha)\Gamma(\beta)}}{\boldsymbol{\Gamma(\alpha+\beta)}}$
- This generalises to: $\boldsymbol{\Delta(\alpha)} = \int_{\boldsymbol{0}}^{\boldsymbol{1}} \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{\alpha_i - 1}}\, \boldsymbol{d\theta} = \frac{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}}{\boldsymbol{\Gamma(A)}}$

$$\boldsymbol{p(c|\alpha)} = \frac{\boldsymbol{C!}}{\prod_{\boldsymbol{i}} \boldsymbol{c_i!}} \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \int \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{c_i + \alpha_i - 1}}\, \boldsymbol{d\theta} = \frac{\boldsymbol{C!}}{\prod_{\boldsymbol{i}} \boldsymbol{c_i!}} \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \frac{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A)}}$$

Posterior Multinomial under Dirichlet prior

- Suppose we observe counts $\boldsymbol{c} = (\boldsymbol{c_1}, \dots, \boldsymbol{c_k})$ from a dice sampled from our factory.
- We would like to predict the most likely parameters $\boldsymbol{\theta}$ for this dice.

$$\boldsymbol{p(\theta|c,\alpha)} = \frac{\boldsymbol{p(c,\theta|\alpha)}}{\boldsymbol{p(c|\alpha)}} = \frac{\boldsymbol{p(c,\theta|\alpha)}}{\int \boldsymbol{p(c,\theta|\alpha)d\theta}} = \frac{\boldsymbol{p(c|\theta)p(\theta|\alpha)}}{\int \boldsymbol{p(c|\theta)p(\theta|\alpha)d\theta}}$$

$$= \frac{\frac{\boldsymbol{C!}}{\prod_i \boldsymbol{c_i!}} \prod_i \boldsymbol{\theta_i^{c_i}} \frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \prod_i \boldsymbol{\theta_i^{\alpha_i - 1}}}{\int \boldsymbol{p(c|\theta)p(\theta|\alpha)d\theta}} = \frac{\frac{\boldsymbol{C!}}{\prod_i \boldsymbol{c_i!}} \frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \prod_i \boldsymbol{\theta_i^{c_i+\alpha_i-1}}}{\int \frac{\boldsymbol{C!}}{\prod_i \boldsymbol{c_i!}} \frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \prod_i \boldsymbol{\theta_i^{c_i+\alpha_i-1}} \boldsymbol{d\theta}}$$

$$= \frac{\prod_i \boldsymbol{\theta_i^{c_i+\alpha_i-1}}}{\frac{\prod_i \boldsymbol{\Gamma(c_i+\alpha_i)}}{\boldsymbol{\Gamma(C+A)}}} = \frac{\boldsymbol{\Gamma(C+A)}}{\prod_i \boldsymbol{\Gamma(c_i+\alpha_i)}} \prod_i \boldsymbol{\theta_i^{c_i+\alpha_i-1}} \sim \boldsymbol{Dir(c_1+\alpha_1, \dots, c_k+\alpha_k)}$$

- So, the shape of the posterior is exactly like that of the prior with counts added into the parameters

Posterior Dirichlet Categorial Conditional Distribution

- Using the Dirichlet Categorical we can compute the probability of observing class $\boldsymbol{k}$ having already observed counts $\boldsymbol{c} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_k})$.
- We will use an indicator variable $\boldsymbol{z} = \boldsymbol{k}$ to indicate that the observed class is $\boldsymbol{k}$:

$$\boldsymbol{p(z = k|c, \alpha)} = \frac{\boldsymbol{p(z = k, c|\alpha)}}{\boldsymbol{p(c|\alpha)}}$$

$$= \frac{\frac{\boldsymbol{\Gamma(A)}}{\boldsymbol{\prod_i \Gamma(\alpha_i)}} \frac{[\boldsymbol{\Gamma(c_k + \alpha_k + 1)}] \boldsymbol{\prod_{i \neq k} \Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A)}}}{\frac{\boldsymbol{\Gamma(A)}}{\boldsymbol{\prod_i \Gamma(\alpha_i)}} \frac{\boldsymbol{\prod_i \Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A - 1)}}}$$

$$= \frac{\boldsymbol{\Gamma(c_k + \alpha_k + 1)}}{\boldsymbol{\Gamma(c_k + \alpha_k)}} \frac{\boldsymbol{\Gamma(C + A - 1)}}{\boldsymbol{\Gamma(C + A)}}$$

$$= \frac{\boldsymbol{c_k + \alpha_k}}{\boldsymbol{C + A - 1}}$$

Not shown here – The multinomial co-efficients will cancel out. So, this distribution is the same for both Categorical and Multnomial

Here: $\boldsymbol{C - 1} = \sum_i \boldsymbol{c_i}$ *So, the total number of items before adding the new* $\boldsymbol{z}$ *is* $\boldsymbol{C - 1}$

- This says that the probability of a new data point being assigned the class $\boldsymbol{k}$ is proportional to $\boldsymbol{c_k + \alpha_k}$, having already observed $\boldsymbol{c}$
- Thus the Dirichlet exhibits the *rich-gets-richer* property

Finite Mixture Model

- Coin/Dice flips are generated from two (or more) coins
- So, the heads/tails/sides are possibly from different coins
- How can we model this?

Finite Mixture Model

- Coin/Dice flips are generated from two (or more) coins
- So, the heads/tails/sides are possibly from different coins
- How can we model this?
- In a finite mixture model we assume that the coin parameter $\boldsymbol{\theta}$ is given by:

$$\boldsymbol{\theta} = \boldsymbol{\pi_1}\boldsymbol{\theta_1} + \boldsymbol{\pi_2}\boldsymbol{\theta_2}$$

  where $\boldsymbol{\pi_1}, \boldsymbol{\pi_2}$ are the *mixture weights* or *proportions*.

- The mixture weights indicate how often each coin has been used. In general:

$$\boldsymbol{\theta} = \sum_{\boldsymbol{i}} \boldsymbol{\pi_i}\boldsymbol{\theta_i}$$

- If we assume, some prior for $\boldsymbol{\theta_i}$ such as the Dirichlet then:

$$\boldsymbol{\theta} = \sum_{\boldsymbol{i}} \boldsymbol{\pi_i}\boldsymbol{p}(\boldsymbol{\theta_i}|\boldsymbol{\alpha})$$

- This model will permit a single coin flip to be generated using multiple coins! So, it is overly general. We will restrict this in the following.

Latent variable models

- We assume that there is a hidden variable **Z** or latent variable that says which mixture component/coin was used to generate the data.
- This makes sense for discrete data such as the coin example since a coin flip cannot be generated from more than one coin.
- Use of hidden variables considerably simplifies the problem formulation and inference
- Given $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$, we assume there is a corresponding latent variable $\boldsymbol{Z} = \{\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_n}\}$ that indicates the mixture component

$$\boldsymbol{p(x_i|\theta) = \sum_k p(x_i|\theta_k) p(z_i = k)}$$

$$\boldsymbol{p(x_i|\theta) = \sum_k p(x_i|\theta_k) \pi_k}$$

where $\boldsymbol{\pi_k}$'s are the *mixture proportions* (or weights)

- Think of $\boldsymbol{z_i}$'s like switches that indicate which coin was used.

Using Dirichlet priors

- To model $\boldsymbol{\pi_k}$'s we can use Dirichlet priors

$$\boldsymbol{p(X, Z, \pi|\theta, \alpha) = p(X, Z|\theta, \pi)p(\pi|\alpha)}$$
$$\boldsymbol{= p(X|Z, \theta)p(Z|\pi)p(\pi|\alpha)}$$

with $\boldsymbol{p(\pi|\alpha) \sim Dir\left(\frac{A}{K}, \ldots, \frac{A}{K}\right)}$ gives distribution over mixture weights/proportions

- Using the Dirichlet-Categorical distribution we can integrate out the mixture proportions $\boldsymbol{\pi}$:

$$\boldsymbol{p(X, Z|\theta, \alpha) = \int p(X, Z, \pi|\theta, \alpha) d\pi}$$
$$\boldsymbol{= p(X|Z, \theta) \int p(Z|\pi)p(\pi|\alpha) d\pi}$$
$$\boldsymbol{= p(X|Z, \theta)\mathrm{p}(\mathrm{Z}|\alpha)}$$
$$\boldsymbol{= p(Z|\alpha) \prod_i p(x_i|Z, \theta)}$$
$$\boldsymbol{= p(Z|\alpha) \prod_i p\left(x_i|\theta_{z_i}\right)}$$

$\boldsymbol{p(Z|\alpha)}$ here would be the Dirichlet-Categorical distribution

Given $\boldsymbol{Z}$ the data points $\boldsymbol{x_i}$'s are independent

Gibbs sampling (review)

- When we have multiple variables to sample e.g. $\boldsymbol{p(X_1, \ldots, X_n)}$
- Then the joint distribution can be estimated by successively estimating:
  - $\boldsymbol{p(X_1 \mid X_2, \ldots, X_n)}$
  - $\boldsymbol{p(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}$
- In the limit, this distribution converges to the joint $\boldsymbol{p(X_1, \ldots, X_n)}$
- More precisely:

$$\boldsymbol{p(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = \frac{p(X_1, \ldots, X_n)}{p(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)} \propto p(X_1, \ldots, X_n)}$$

Gibbs sampling for finite mixtures

- The mixture proportions are determined by $\boldsymbol{Z}$. In particular, the count $\boldsymbol{c_k}$ of the number of data points in class $\boldsymbol{k}$ is given by:
  - $\boldsymbol{c_k = |\{z_i = k| \, z_i \in Z\}|}$
- We need to estimate $\boldsymbol{p(x_i, z_i | X_{-i}, Z_{-i}, \theta, \alpha)}$:

$$\boldsymbol{p(x_i, z_i | X_{-i}, Z_{-i}, \theta, \alpha) p(X_{-i}, Z_{-i} | \theta, \alpha) = p(X, \mathbf{Z} | \theta, \alpha)}$$

- Thus: $\boldsymbol{p(x_i, z_i | X_{-i}, Z_{-i}, \theta, \alpha) = \dfrac{p(X, \mathbf{Z} | \theta, \alpha)}{p(X_{-i}, Z_{-i} | \theta, \alpha)}}$

$$\boldsymbol{= \frac{p(X | \, \mathbf{Z}, \theta)}{p(X_{-i} | Z_{-i}, \theta)} \frac{p(\mathbf{Z} | \alpha)}{p(Z_{-i} | \alpha)}}$$

$$\boldsymbol{= p(x_i | z_i, \theta) \frac{p(\mathbf{Z} | \alpha)}{p(Z_{-i} | \alpha)}}$$

$$\boldsymbol{= p\left(x_i | \theta_{z_i}\right) \frac{p(\mathbf{Z} | \alpha)}{p(Z_{-i} | \alpha)}}$$

$$\boldsymbol{= p\left(x_i | \theta_{z_i}\right) p(z_i | Z_{-i}, \alpha)}$$

Since:

$$\boldsymbol{p(\mathbf{z_i}, \mathbf{Z_{-i}} | \alpha) = p(\mathbf{Z} | \alpha)}$$
$$\boldsymbol{= p(z_i | Z_{-i}, \alpha) p(\mathbf{Z_{-i}} | \alpha)}$$

Posterior Dirichlet Categorial Conditional Distribution

- Using the Dirichlet Categorical we can compute the probability of observing class $\boldsymbol{k}$ having already observed counts $\boldsymbol{c} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_k})$:

$$\boldsymbol{p(z = k|c, \alpha)} = \frac{\boldsymbol{p(z = k, c|\alpha)}}{\boldsymbol{p(c|\alpha)}}$$

$$= \frac{\frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \frac{[\boldsymbol{\Gamma(c_k + \alpha_k + 1)}] \prod_{i \neq k} \boldsymbol{\Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A)}}}{\frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \frac{\prod_i \boldsymbol{\Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A - 1)}}}$$

$$= \frac{\boldsymbol{\Gamma(c_k + \alpha_k + 1)}}{\boldsymbol{\Gamma(c_k + \alpha_k)}} \frac{\boldsymbol{\Gamma(C + A - 1)}}{\boldsymbol{\Gamma(C + A)}}$$

$$= \frac{\boldsymbol{c_k + \alpha_k}}{\boldsymbol{C + A - 1}}$$

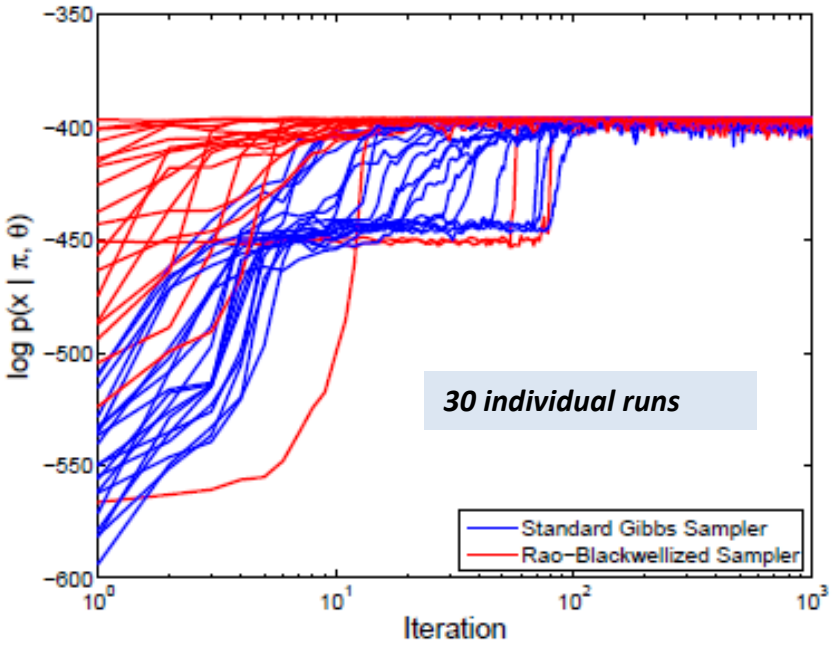
Not shown here –
The multinomial co-efficients will cancel out. So, this distribution is the same for both Categorical and Multnomial

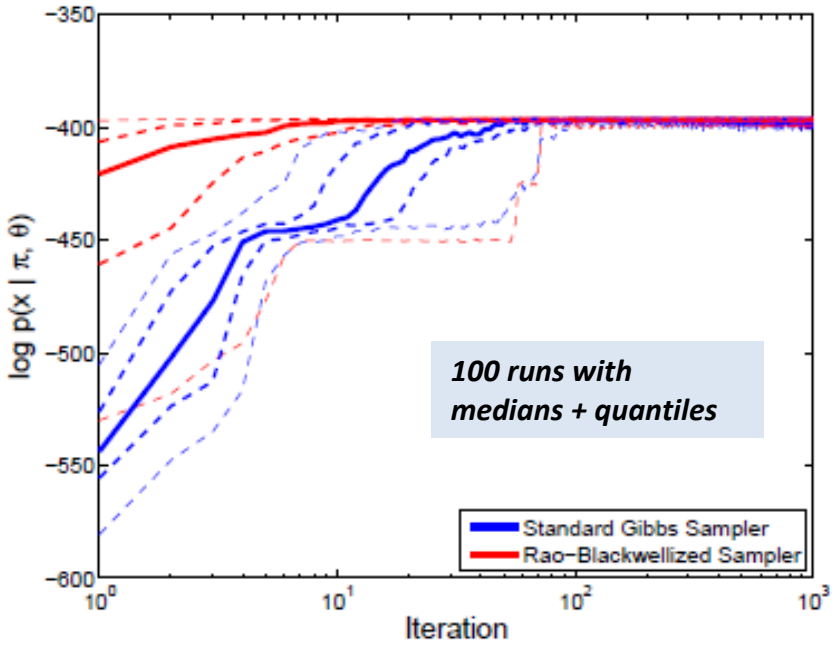
Here: $\boldsymbol{C - 1} = \sum_i \boldsymbol{c_i}$
So, the total number of items before adding the new ***z*** *is* $\boldsymbol{C - 1}$

- This says that the probability of a new data point being assigned the class $\boldsymbol{k}$ is proportional to $\boldsymbol{c_k + \alpha_k}$, having already observed $\boldsymbol{c}$
- Thus the Dirichlet exhibits the *rich-gets-richer* property

Collapsed Gibbs Sampling

- Using the Dirichlet Multinomial to estimate the mixture proportions is a form of ***collapsed sampling*** or ***Rao-Blackwellized Gibbs sampling***
- Collapsed sampling results in Markov chains that converge much quicker and gets trapped in local optimal less often





From [Sudderth]

Finite Mixtures (θ known)

- In the simplest case, the (class density) parameters $\boldsymbol{\theta}$ are known i.e. you know the bias of each coin
- So, the task is to determine the $\boldsymbol{z_i}$'s i.e. to determine which coin was used for each $\boldsymbol{x_i}$

Randomly assign $\boldsymbol{z_i = k}$ with uniform probability

Repeat for each $\boldsymbol{i}$:

Remove data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$

For each cluster $\boldsymbol{k}$ compute $\boldsymbol{p_k(x_i|\,\theta_k)}$

Add data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$ back by sampling $\boldsymbol{z_i = k}$ using:

$$\boldsymbol{z_i = k \sim \frac{c_k + A/K}{N - 1 + A} p_k(x_i|\,\theta_k)}$$

Mixture proportions can be estimated at the end

Finite Mixtures (θ unknown)

- In the general case, the (class density) parameters $\boldsymbol{\theta}$ are unknown i.e. you do not know the bias of each coin
- So, the task is to determine both
  - the $\boldsymbol{z_i}$'s i.e. to determine which coin was used for each $\boldsymbol{x_i}$ and
  - the class density parameters $\boldsymbol{\theta}$ i.e. the bias for each coin
- In this case, the joint density we need to estimate:

$$\begin{aligned} \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \boldsymbol{p}(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{\beta}) \\ &= \boldsymbol{p}(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{Z}|\boldsymbol{\alpha})\boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{\beta}) \end{aligned}$$

- Following the derivation from page 34, gives:

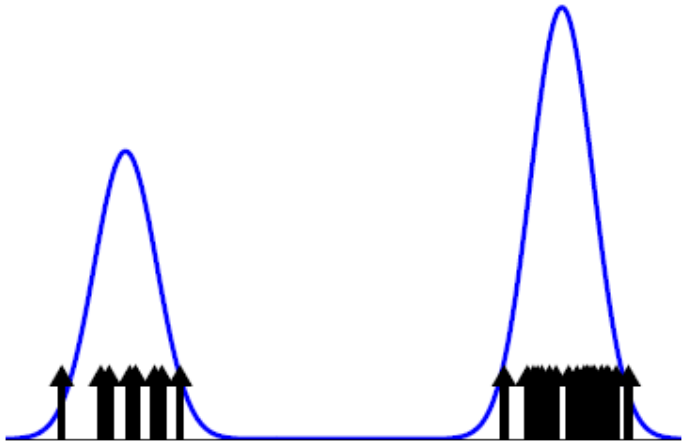
$$\boldsymbol{p}(\boldsymbol{x_i}, \boldsymbol{z_i}|\boldsymbol{X_{-i}}, \boldsymbol{Z_{-i}}, \boldsymbol{\theta}, \boldsymbol{\alpha}) = \boldsymbol{p}\left(\boldsymbol{x_i}|\boldsymbol{\theta_{z_i}}\right) \boldsymbol{p}(\boldsymbol{z_i}|\boldsymbol{Z_{-i}}, \boldsymbol{\alpha})$$

- To sample $\boldsymbol{\theta}$ we need to use the posterior distribution:

$$\boldsymbol{p}(\boldsymbol{\theta_k}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}) = \boldsymbol{p}(\boldsymbol{\theta_k}|\{\boldsymbol{x_i}|\boldsymbol{z_i} = \boldsymbol{k} \in \boldsymbol{Z}\}, \boldsymbol{\beta_k})$$

- For the Multinomial with Dirichlet prior see page 27

Mixture of Univariate Gaussians



- Here the aim is to estimate the Gaussian parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ and the mixture proportions using the finite mixture model

Mixture of Univariate Gaussians

- Randomly assign $\boldsymbol{z_i = k}$ with uniform probability
- Repeat for each $\boldsymbol{i}$:
  - Remove data point $(\boldsymbol{x_i, z_i})$
  - Estimate Gaussian parameters $\boldsymbol{\theta_k = (\mu_k, \sigma_k)}$ for each cluster from current cluster assignments e.g. using MLE. For MLE, we can ignore the prior $\boldsymbol{\beta}$

    $$\boldsymbol{p(\mu_k, \sigma_k | X, Z) = p(\mu_k, \sigma_k | \{x_i | z_i = k \in Z\})}$$

    For each cluster $\boldsymbol{k}$ compute $\boldsymbol{p(x_i | \mu_k, \sigma_k)}$
  - Sample $\boldsymbol{z_i}$ using:

    $$\boldsymbol{z_i = k \sim \frac{c_k + A/K}{C - 1 + A} p(x_i | \mu_k, \sigma_k)}$$
- Mixture proportions can be estimated at the end

Mixture of Multinomials

- For each data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$ randomly assign $\boldsymbol{z_i = k}$ with uniform probability
- Repeat for each $\boldsymbol{i}$:
  - Remove data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$
  - Sample Multinomial parameters $\boldsymbol{\theta_k}$ for each cluster from current cluster assignments $\boldsymbol{X_k} = \{\boldsymbol{x_i} \in \boldsymbol{X} \,|\boldsymbol{z_i} = \boldsymbol{k}\}$ using the posterior Multinomial distribution:

$$\boldsymbol{p(\theta|c,\beta)} = \frac{\boldsymbol{\Gamma(C+B)}}{\boldsymbol{\prod_i \Gamma(c_i+\beta_i)}} \prod_{\boldsymbol{i}} \boldsymbol{\theta_i^{c_i+\beta_i-1}}$$

  - For each cluster $\boldsymbol{k}$ compute $\boldsymbol{p_k(x_i|\,\theta_k)}$
  - Add data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$ back by sampling $\boldsymbol{z_i = k}$ using:

$$\boldsymbol{z_i = k} \sim \frac{\boldsymbol{c_k + \alpha_k}}{\boldsymbol{C - 1 + A}} \boldsymbol{p_k(x_i|\,\theta_k)}$$

  where $\boldsymbol{p_k(x_i|\,\theta_k)}$ is the multinomial distribution for class $\boldsymbol{k}$
- Mixture proportions can be estimated at the end

Collapsed Sampler for Mixture of Multinomials

- Like we did for the mixture proportions, the within-class multinomial parameters $\boldsymbol{\theta_k}$ can be integrated out using the Dirichlet-Multinomial
- Thus instead of

$$\boldsymbol{p}\left(\boldsymbol{x_i}|\boldsymbol{\theta}_{\boldsymbol{z_i}}\right)$$

we need to compute the factor:

$$\frac{\boldsymbol{p}(\boldsymbol{X}|\,\mathbf{Z},\boldsymbol{\beta})}{\boldsymbol{p}(\boldsymbol{X}_{-\boldsymbol{i}}|\boldsymbol{Z}_{-\boldsymbol{i}},\boldsymbol{\beta})}$$

- If we assume that for $(\boldsymbol{x_i}, \boldsymbol{z_i})$ the target destination class is $\boldsymbol{k}$ i.e. $\boldsymbol{z_i} = \boldsymbol{k}$:

$$\frac{\boldsymbol{p}(\boldsymbol{X}|\,\mathbf{Z},\boldsymbol{\beta})}{\boldsymbol{p}(\boldsymbol{X}_{-\boldsymbol{i}}|\boldsymbol{Z}_{-\boldsymbol{i}},\boldsymbol{\beta})} = \frac{\prod_{\boldsymbol{k}}\boldsymbol{p}(\boldsymbol{X}_{\boldsymbol{k}}|\,\boldsymbol{\beta})}{\prod_{\boldsymbol{k}}\boldsymbol{p}\left(\boldsymbol{X}_{\boldsymbol{k},-\boldsymbol{i}}|\boldsymbol{\beta}\right)} = \frac{\boldsymbol{p}(\boldsymbol{X}_{\boldsymbol{k}}|\boldsymbol{\beta})}{\boldsymbol{p}\left(\boldsymbol{X}_{\boldsymbol{k},-\boldsymbol{i}}|\boldsymbol{\beta}\right)}$$

- In the above, $\boldsymbol{X_k} = \{\boldsymbol{x_i} \in \boldsymbol{X}\,|\boldsymbol{z_i} = \boldsymbol{k}\}$ (as discussed in previous slides)
- It is worth remembering that each $\boldsymbol{x_i}$ is a side of the dice
- Thus using the Dirichlet prior $\boldsymbol{\beta}$ in the Dirichlet-Multinomial distribution we can estimate the within-class likelihood $\boldsymbol{p}(\boldsymbol{X_k}|\boldsymbol{\beta})$

$$\frac{p(X_k|\beta)}{p(X_{k,-i}|\beta)} = \frac{\frac{\Gamma(B)}{\prod_j \Gamma(\beta_j)} \frac{\left[\Gamma\left(c_{x_i}^k + \beta_{x_i} + 1\right)\right] \prod_{j \neq x_i} \Gamma(c_j + \beta_j)}{\Gamma(C^j + B)}}{\frac{\Gamma(B)}{\prod_j \Gamma(\beta_j)} \frac{\prod_j \Gamma(c_j + \beta_j)}{\Gamma(C^j + B - 1)}}$$

$$= \frac{\Gamma\left(c_{x_i}^k + \beta_{x_i} + 1\right)}{\Gamma\left(c_{x_i}^k + \beta_{x_i}\right)} \frac{\Gamma\left(C^k + B - 1\right)}{\Gamma(C^k + B)} = \frac{c_{x_i}^k + \beta_{x_i}}{C^k + B - 1}$$

- j in the above, iterates over the total number of sides of the dice
- c_{x_i} is the count of the number of observations having the same side as x_i in class k

Fully collapsed sampler

- Randomly assign $\boldsymbol{z_i = k}$ with uniform probability
- Repeat for each $\boldsymbol{i}$:
  - Remove data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$
  - For each cluster $\boldsymbol{k}$ compute $\frac{\boldsymbol{c_{x_i}^k + \beta_{x_i}}}{\boldsymbol{C^k + B - 1}}$
  - Add data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$ back by sampling $\boldsymbol{z_i = k}$ using:

$$\boldsymbol{z_i = k \sim \frac{c_k + A/K}{N - 1 + A}\left(\frac{c_{x_i}^k + \beta_{x_i}}{C^k + B - 1}\right)}$$

- Mixture proportions can be estimated at the end

In the above $\boldsymbol{c_{x_i}^k}$ is the count of the number of observations having the same side as $\boldsymbol{x_i}$ in class $\boldsymbol{k}$

Topic models

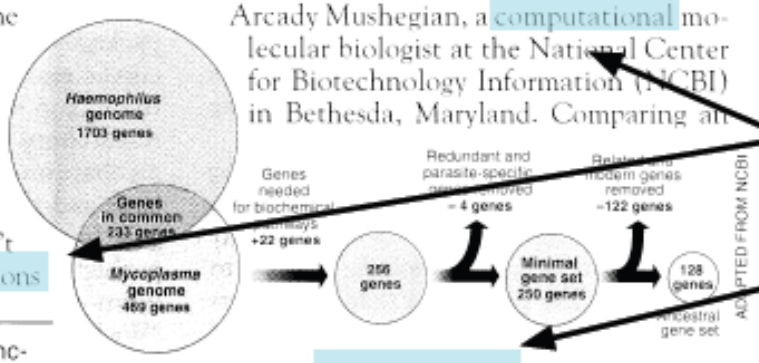
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

***From* [Blei]**

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson [illegible] Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



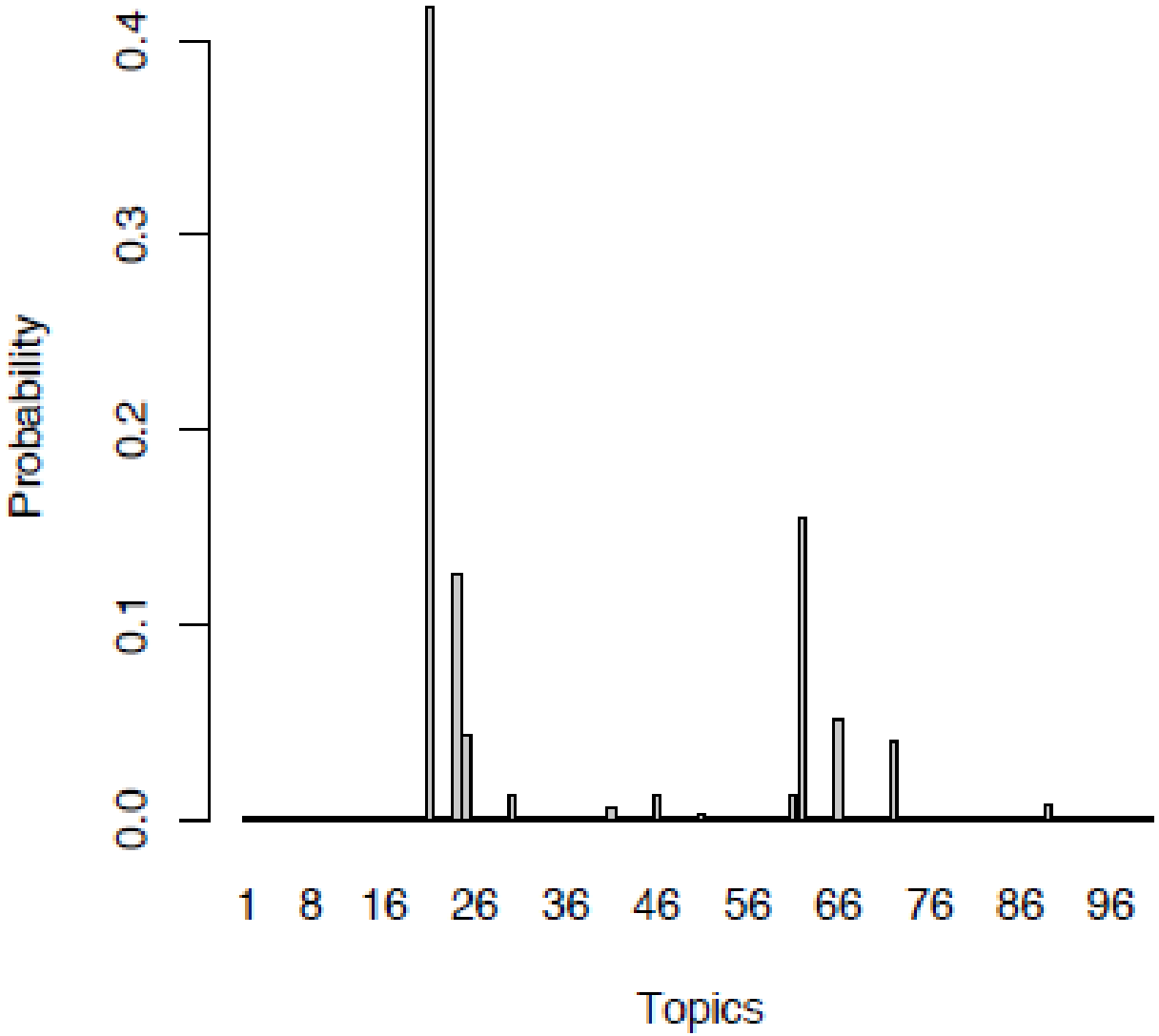
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

***From* [Blei]**

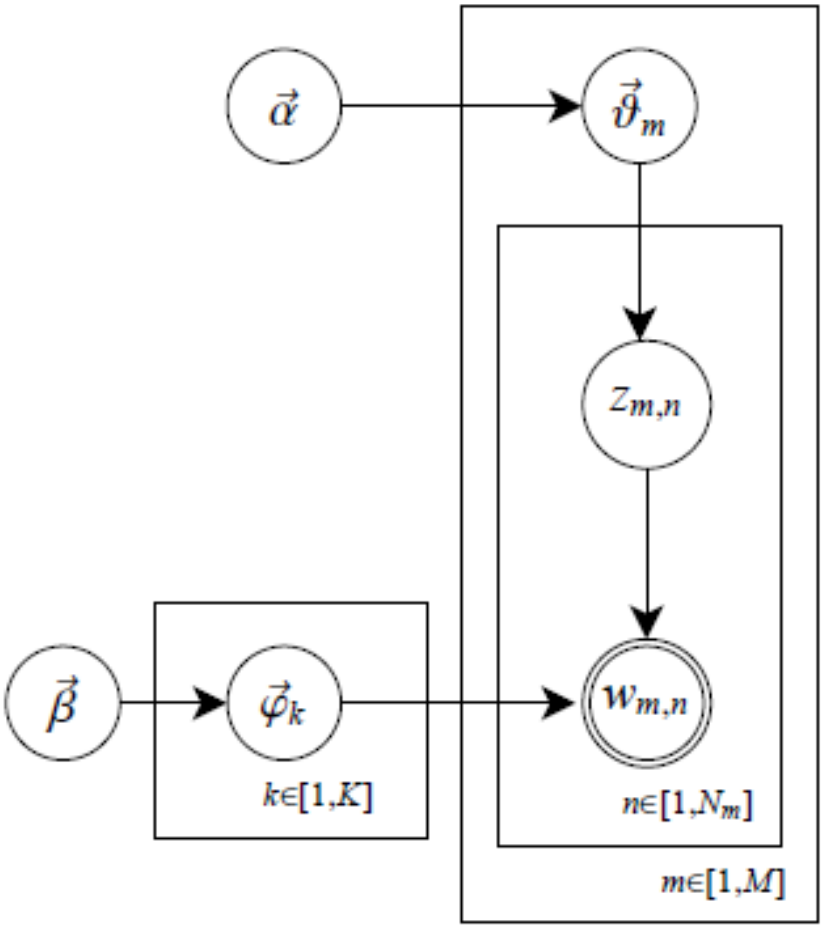
Document-topic distribution



Latent Dirichlet Allocation

- LDA is a hidden variable generative model for text documents
- The documents are modelled as bag-of-words or a Categorical distribution over words
- So, words are the sides of a dice and the number of times seen in a document is its count.
- However, each document is generated by a mixture of k Categorical distributions.
- But these distributions are shared across all documents
- So, all documents are generated by the same set of dices.
- The only difference between two documents are their mixture proportions

LDA plate diagram



- Use of Dirichlet Multinomials means that the multinomial parameters are integrated out.

Modeling the document-topic distribution

- Assume that we have $\boldsymbol{K}$ topics and $\boldsymbol{D}$ documents
- Let $\boldsymbol{c_d} = (\boldsymbol{c_{d,1}}, \dots, \boldsymbol{c_{d,K}})$ be a $\boldsymbol{K}$-dimensional vector such that $\boldsymbol{c_{d,k}}$ is the count of words with topic $\boldsymbol{k}$ in document $\boldsymbol{d}$
- Let $\boldsymbol{C_d} = \sum_{\boldsymbol{k}} \boldsymbol{c_{d,k}}$ is the count of words in document $\boldsymbol{d}$
- Then, the *document-to-topic distribution* can be modelled using the Dirichlet Multinomial distribution:

$$\boldsymbol{p(c_d|\alpha)} = \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{k}}^{\boldsymbol{K}} \boldsymbol{\Gamma(\alpha_k)}} \frac{\prod_{\boldsymbol{k}}^{\boldsymbol{K}} \boldsymbol{\Gamma}\left(\boldsymbol{c_{d,k}} + \boldsymbol{\alpha_k}\right)}{\boldsymbol{\Gamma(C_d + A)}}$$

- Let $\boldsymbol{Z} = (\boldsymbol{z_1}, \dots, \boldsymbol{z_D})$ be a $\boldsymbol{D}$-dimensional vector such that each $\boldsymbol{z_i} = (\boldsymbol{z_{i,1}}, \dots, \boldsymbol{z_{i,|z_i|}})$ is a vector of topic indicator variables for document $\boldsymbol{i}$
- The topic assignments within each document are independent of the topic assignments in other documents.

$$\boldsymbol{p(Z|\alpha)} = \prod_{\boldsymbol{d}}^{\boldsymbol{D}} \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{k}}^{\boldsymbol{K}} \boldsymbol{\Gamma(\alpha_k)}} \frac{\prod_{\boldsymbol{k}}^{\boldsymbol{K}} \boldsymbol{\Gamma}\left(\boldsymbol{c_{d,k}} + \boldsymbol{\alpha_k}\right)}{\boldsymbol{\Gamma(C_d + A)}}$$

Modelling the topic-word distribution

- Let $\boldsymbol{n_k} = (\boldsymbol{n_{k,1}}, \ldots, \boldsymbol{n_{k,V}})$ be a **V**-dimensional vector such that $\boldsymbol{n_{k,v}}$ is the count of word **v** in topic **k** across the ***whole*** document collection
- Let $\boldsymbol{N_k} = \sum_{\boldsymbol{v}}^{\boldsymbol{V}} \boldsymbol{n_{k,v}}$ is the total number of words in topic $\boldsymbol{k}$
- Then we can use the Dirichlet Multinomial (again) to model the *topic-to-word* distribution:

$$\boldsymbol{p(n_k|Z,\beta)} = \frac{\boldsymbol{\Gamma(B)}}{\prod_{\boldsymbol{v}}^{\boldsymbol{V}} \boldsymbol{\Gamma(\beta_v)}} \frac{\prod_{\boldsymbol{v}}^{\boldsymbol{V}} \boldsymbol{\Gamma(n_{k,v} + \beta_v)}}{\boldsymbol{\Gamma(N_k + B)}}$$

- We need the $\boldsymbol{Z}$ to collect all words belonging to a particular topic
- The topic-word distribution for one topic is independent of the same for another topic
- Let $\boldsymbol{W} = (\boldsymbol{w_1}, \ldots, \boldsymbol{w_D})$ be a $\boldsymbol{D}$-dimensional vector such that each $\boldsymbol{w_i} = (\boldsymbol{w_{i,1}}, \ldots, \boldsymbol{w_{i,|z_i|}})$ is a vector of words for document $\boldsymbol{i}$. Then:

$$\boldsymbol{p(W|Z,\beta)} = \prod_{\boldsymbol{k}}^{\boldsymbol{K}} \frac{\boldsymbol{\Gamma(B)}}{\prod_{\boldsymbol{v}}^{\boldsymbol{V}} \boldsymbol{\Gamma(\beta_v)}} \frac{\prod_{\boldsymbol{v}}^{\boldsymbol{V}} \boldsymbol{\Gamma(n_{k,v} + \beta_v)}}{\boldsymbol{\Gamma(N_k + B)}}$$

- Thus, the joint distribution becomes:

$$p(W, Z|\alpha, \beta) = p(W|Z, \beta)\, p(Z|\alpha)$$

$$= \left[\prod_{k}^{K} \frac{\Gamma(B)}{\prod_{v}^{V}\Gamma(\beta_v)} \frac{\prod_{v}^{V}\Gamma(n_{k,v}+\beta_v)}{\Gamma(N_k+B)}\right]\left[\prod_{d}^{D} \frac{\Gamma(A)}{\prod_{k}^{K}\Gamma(\alpha_k)} \frac{\prod_{k}^{K}\Gamma(c_{d,k}+\alpha_k)}{\Gamma(C_d+A)}\right]$$

- To do Gibbs sampling, we need to sample $z_i = k$ where $i = (m, n)$

$$p(w_i, z_i = k|W_{-i}, Z_{-i}, \alpha, \beta) = \frac{p(W, Z|\alpha, \beta)}{p(W_{-i}, Z_{-i}|\alpha, \beta)} = \frac{p(W|Z, z_i = k, \beta)}{p(W_{-i}|Z_{-i}, \beta)} \frac{p(Z, z_i = k|\alpha)}{p(Z_{-i}|\alpha)}$$

$$\frac{p(Z, z_i = k, i = (m, n)|\alpha)}{p(Z_{-i}|\alpha)} =$$

$$\frac{\left[\frac{\Gamma(A)}{\prod_{j}^{K}\Gamma(\alpha_j)} \frac{\left[\Gamma(c_{m,k}+\alpha_k+1)\right]\prod_{j\neq k}^{K}\Gamma(c_{m,j}+\alpha_j)}{\Gamma(C_m+A)}\right]\prod_{d\neq m}^{D} \frac{\Gamma(A)}{\prod_{j}^{K}\Gamma(\alpha_j)} \frac{\prod_{j}^{K}\Gamma(c_{d,j}+\alpha_j)}{\Gamma(C_d+A)}}{\left[\frac{\Gamma(A)}{\prod_{j}^{K}\Gamma(\alpha_j)} \frac{\left[\Gamma(c_{m,k}+\alpha_k)\right]\prod_{j\neq k}^{K}\Gamma(c_{m,j}+\alpha_j)}{\Gamma(C_m+A-1)}\right]\prod_{d\neq m}^{D} \frac{\Gamma(A)}{\prod_{j}^{K}\Gamma(\alpha_j)} \frac{\prod_{j}^{K}\Gamma(c_{d,j}+\alpha_j)}{\Gamma(C_d+A)}}$$

$$= \frac{\Gamma(c_{m,k}+\alpha_k+1)}{\Gamma(C_m+A)} \frac{\Gamma(C_m+A-1)}{\Gamma(c_{m,k}+\alpha_k)} = \frac{c_{m,k}+\alpha_k}{C_m+A-1}$$

$$\frac{p(W, z_i = k, i = (m,n)|Z, \beta)}{p(W_{-i}|Z_{-i}, \beta)} =$$

$$\frac{\left[\frac{\Gamma(B)}{\prod_v^V \Gamma(\beta_v)} \frac{\left[\Gamma(n_{k,i} + \beta_i + 1)\right] \prod_{v \neq i}^V \Gamma(n_{k,v} + \beta_v)}{\Gamma(N_k + B)}\right] \prod_{j \neq k}^K \frac{\Gamma(B)}{\prod_v^V \Gamma(\beta_v)} \frac{\prod_{v \neq i}^V \Gamma(n_{j,v} + \beta_v)}{\Gamma(N_j + B)}}{\left[\frac{\Gamma(B)}{\prod_v^V \Gamma(\beta_v)} \frac{\left[\Gamma(n_{k,i} + \beta_i)\right] \prod_{v \neq i}^V \Gamma(n_{k,v} + \beta_v)}{\Gamma(N_k + B - 1)}\right] \prod_{j \neq k}^K \frac{\Gamma(B)}{\prod_v^V \Gamma(\beta_v)} \frac{\prod_v^V \Gamma(n_{j,v} + \beta_v)}{\Gamma(N_j + B)}}$$

$$= \frac{\Gamma(n_{k,i} + \beta_i + 1)}{\Gamma(n_{k,i} + \beta_i)} \frac{\Gamma(N_k + B - 1)}{\Gamma(N_k + B)}$$

$$= \frac{n_{k,i} + \beta_i}{N_k + B - 1}$$

- ***Thus:***

$$p(w_i, z_i = k, i = (m,n)|W_{-i}, Z_{-i}, \alpha, \beta) = \frac{c_{m,k} + \alpha_k}{C_m + A - 1} \quad \frac{n_{k,i} + \beta_i}{N_k + B - 1}$$

Note: Here $v \neq i$ means that the v^{th} word in the vocabulary is not the same as the i^{th} word in the document collection. The i^{th} word in the document collection being the same as the n^{th} word in the m^{th} document.

LDA algorithm sketch

- Randomly assign $z_i = k$ where $i = (m, n)$ with uniform probability. Remember i is also a pointer to a word
- Update variables c_m, n_k, C_m, N_k appropriately
- c is $D \times K$ int matrix; n is $K \times V$ int matrix
- C is D-dim vector of int; N is K-dim vector of int
- Repeat for each $i = (m, n)$:
  - Remove data point (x_i, z_i)
  - Sample $z_i = k$ using:

$$p(w_i, z_i = k | W, Z_{-i}, \alpha, \beta) = \frac{c_{m,k} + \alpha_k}{C_m + A - 1} \quad \frac{n_{k,i} + \beta_i}{N_k + B - 1}$$

  - Update variables c_m, n_k, C_m, N_k appropriately
- Mixture proportions can be estimated at the end

Infinite Mixture Models

Motivation

- Key problem with finite mixtures is that the number of mixture components (**K**) needs to be given in advance
- For large datasets/problems, it is often impossible to figure out what an optimal value for **K** should be
- The Dirichlet Process (**DP**) is an extension of the Dirichlet distribution to the infinite case

Dirichlet Categorical (revision)

- Like for the Beta-Binomial distribution we can integrate out the Categorical parameters
- Here, we are modelling the case where we need to predict the outcome $\boldsymbol{c} = (\boldsymbol{c_1}, \dots, \boldsymbol{c_K})$ of a dice throw from a dice sampled from a factory with Dirichlet parameter $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k})$

$$\boldsymbol{p(c|\alpha)} = \int \boldsymbol{p(c, \theta|\alpha)}\, \boldsymbol{d\theta} = \int \boldsymbol{p(c|\theta)}\, \boldsymbol{p(\theta|\alpha)}\, \boldsymbol{d\theta}$$

$$= \int \left[\prod_{\boldsymbol{i}} \boldsymbol{\theta}^{\boldsymbol{c_i}}\right] \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{\alpha_i - 1}}\, \boldsymbol{d\theta}$$

$$= \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \int \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{c_i + \alpha_i - 1}}\, \boldsymbol{d\theta}$$

- From earlier: $\boldsymbol{B(\alpha, \beta)} = \int_{\boldsymbol{0}}^{\boldsymbol{1}} \boldsymbol{\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta} = \frac{\boldsymbol{\Gamma(\alpha)\Gamma(\beta)}}{\boldsymbol{\Gamma(\alpha+\beta)}}$
- This generalises to: $\boldsymbol{\Delta(\alpha)} = \int_{\boldsymbol{0}}^{\boldsymbol{1}} \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{\alpha_i - 1}}\, \boldsymbol{d\theta} = \frac{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}}{\boldsymbol{\Gamma(A)}}$

$$\boldsymbol{p(c|\alpha)} = \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \int \prod_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{c_i + \alpha_i - 1}}\, \boldsymbol{d\theta} = \frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \frac{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A)}}$$

Posterior Conditional Distribution (revision)

- Having observed counts $\boldsymbol{c} = (\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_K})$, the probability of observing class $\boldsymbol{k}$ is given by the posterior conditional distribution
- Using the Dirichlet Categorical we can compute the probability of observing class $\boldsymbol{k}$ having already observed $\boldsymbol{c}$

$$\boldsymbol{p(z = k|c, \alpha)} = \frac{\boldsymbol{p(z = k, c|\alpha)}}{\boldsymbol{p(c|\alpha)}}$$

$$= \frac{\frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \frac{[\boldsymbol{\Gamma(c_k + \alpha_k + 1)}] \prod_{\boldsymbol{i \neq k}} \boldsymbol{\Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A)}}}{\frac{\boldsymbol{\Gamma(A)}}{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(\alpha_i)}} \frac{\prod_{\boldsymbol{i}} \boldsymbol{\Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A - 1)}}}$$

$$= \frac{\boldsymbol{\Gamma(c_k + \alpha_k + 1)}}{\boldsymbol{\Gamma(c_k + \alpha_k - 1)}} \frac{\boldsymbol{\Gamma(C + A - 1)}}{\boldsymbol{\Gamma(C + A)}}$$

$$= \frac{\boldsymbol{c_k + \alpha_k}}{\boldsymbol{C + A - 1}}$$

Here: $\boldsymbol{C - 1} = \sum_{\boldsymbol{i}} \boldsymbol{c_i}$
So, the total number of items before adding the new $\boldsymbol{z}$ *is* $\boldsymbol{C - 1}$

- This says that the probability of a new data point being assigned the class $\boldsymbol{k}$ is proportional to $\boldsymbol{c_k + \alpha_k}$, having already observed $\boldsymbol{c}$
- Thus the Dirichlet exhibits the *rich-gets-richer* property

The Chinese Restaurant Process

- Firstly, lets assume that we will use a symmetric Dirichlet prior i.e. assume that $\boldsymbol{\alpha} = \left(\frac{\boldsymbol{A}}{\boldsymbol{K}}, \ldots, \frac{\boldsymbol{A}}{\boldsymbol{K}}\right)$ i.e. each $\boldsymbol{\alpha_i} = \frac{\boldsymbol{A}}{\boldsymbol{K}}$
- For the Dirichlet Categorical, the posterior conditional of observing class $\boldsymbol{k}$ having already observed counts $\boldsymbol{c}$ is:

$$\boldsymbol{p(z = k|c, \alpha) = \frac{c_k + \alpha_k}{C + A - 1}}$$

- Now if we let $\boldsymbol{\alpha_k = \frac{A}{K}}$ and re-parameterise (using $\boldsymbol{A}$) then:

$$\boldsymbol{p(z = k|c, A) = \frac{c_k + \frac{A}{K}}{C + A - 1}}$$

- Now we pose the question: ***What happens when*** $\boldsymbol{K \to \infty}$?

$$\boldsymbol{\lim_{K \to \infty} p(z = k|c, A) = \lim_{K \to \infty} \frac{c_k + \frac{A}{K}}{C + A - 1} = \frac{c_k}{C + A - 1}}$$

- So:

$$\boldsymbol{p(z = k|c, A) = \frac{c_k}{C + A - 1}}$$

for a class $\boldsymbol{k}$ that is already present i.e. $\boldsymbol{c_k > 0}$

- Since, this is a distribution:

$$\boldsymbol{\sum_{i}^{K} p(z = k|c, A) = 1}$$

- But we find that:

$$\boldsymbol{\sum_{i}^{K} p(z = k|c, A) = \sum_{i}^{K} \frac{c_i}{C + A - 1} = \frac{C - 1}{C + A - 1} < 1}$$

$$\boldsymbol{1 - \frac{C - 1}{C + A - 1} = \frac{A}{C + A - 1}}$$

- This remaining probability mass is for the remaining unseen classes.

- The ***Chinese Restaurant Process (CRP)*** is defined by the conditional distribution:

$$p(z = k|c, A) = \frac{c_k}{C+A-1} \quad \text{if } k \text{ is an existing class}$$

$$p(z = K+1|c, A) = \frac{A}{C+A-1} \text{ for a new class } K+1$$

- More conventionally, it is written as:

$$p(z = k|c, \alpha) = \frac{c_k}{C+\alpha-1} \quad \text{if } k \text{ is an existing class}$$

$$p(z = K+1|c, \alpha) = \frac{\alpha}{C+\alpha-1} \text{ for new } K+1$$

Here: α *is no longer a vector but just a number.*

- So, the CRP is defined as:

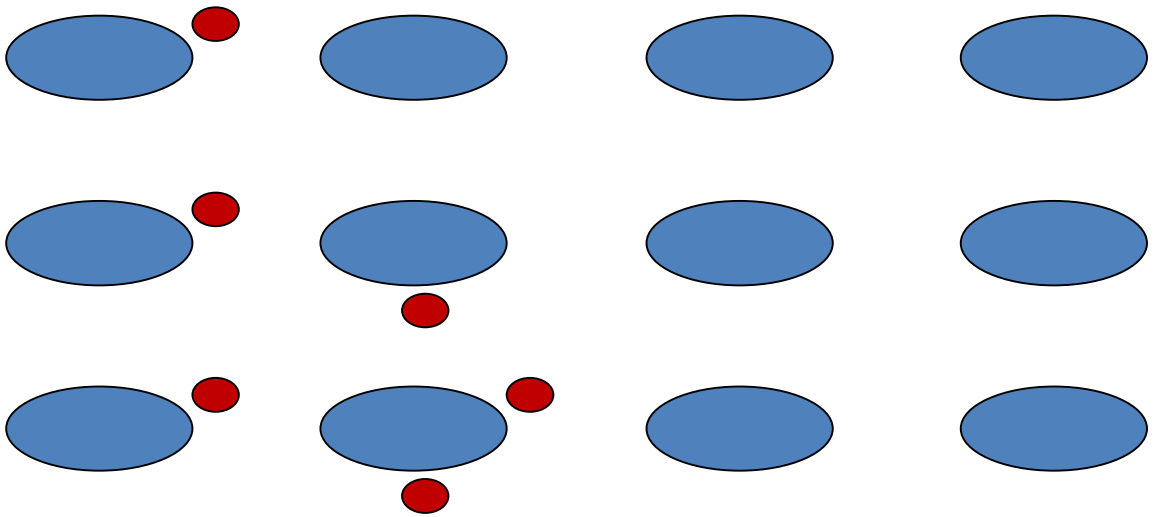
$$p(z = k|c, \alpha) = CRP(z = k|c, \alpha) = \frac{c_k}{C+\alpha-1}$$ for existing class k

$$p(z = K + 1|c, \alpha) = CRP(z = k + 1|c, \alpha) = \frac{\alpha}{C+\alpha-1}$$ for new class

- α is the ***concentration parameter*** that controls how the mass is assigned to existing clusters
- Smaller values of α concentrate the probability mass to **existing** clusters
- Larger values of α concentrate the probability mass to **new** clusters

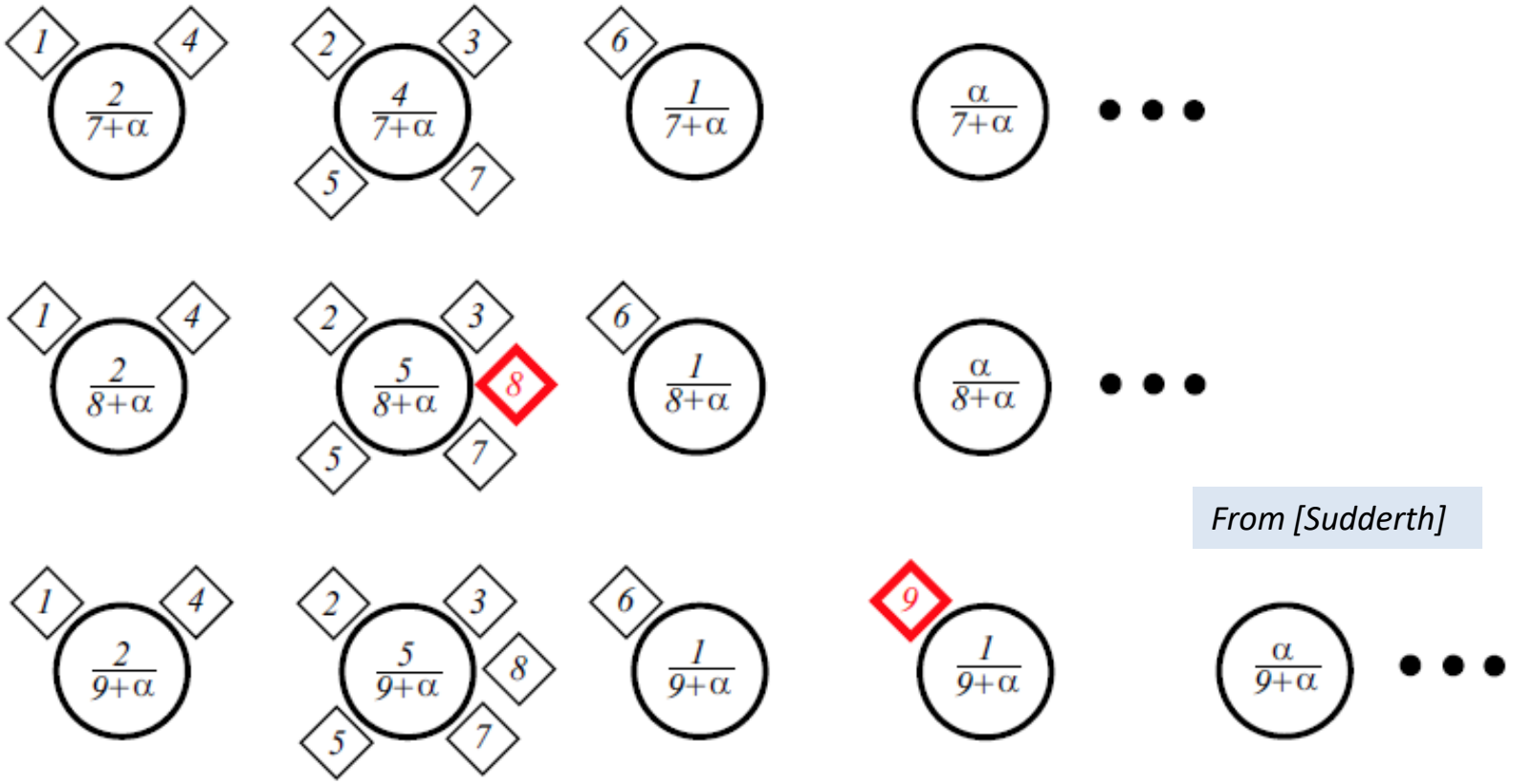
- The CRP is the infinite extension of the Dirichlet Categorical posterior conditional/predictive distribution.
- Since, the mixture proportions π_k have been integrated out, the CRP is a distribution over counts.

The Chinese Restaurant



- The *first* customer occupies the first table with prob. $\frac{\alpha}{1+\alpha-1} = 1$
- The *second* customer either occupies the first table with prob. $\frac{1}{2+\alpha-1} = \frac{1}{1+\alpha}$ or can start a new table with prob. $\frac{\alpha}{2+\alpha-1} = \frac{\alpha}{1+\alpha}$
- The N^{th} customer occupies a table k with prob. $\frac{c_k}{N+\alpha-1}$ or can start a new table with prob. $\frac{\alpha}{N+\alpha-1}$

Another illustration of CRP



From [Sudderth]

- Middle row : 8th customer joins the most popular table
- Last row : 9th customer starts a new table

The Polya urn scheme

An equivalent representation can be obtained by considering the following scheme

- An urn contains different coloured balls
- With probability $\frac{C-1}{C+\alpha-1}$:
  - Pick a ball from the urn (with uniform probability)
  - Put back the ball into the urn with an **extra** ball of the same colour
- With probability $\frac{\alpha}{C+\alpha-1}$:
  - Draw a new ball of different colour from somewhere else and add this into the urn

The Polya urn scheme is equivalent to the Chinese Restaurant metaphor.

The rich-gets-richer property is clearly evident.

Polya urn representation

- Equivalently, the CRP can be defined as:

$$\boldsymbol{p(z = k|Z, \alpha) = CRP(z = k|c, \alpha) = \sum_i \frac{I(z_i = k)}{C + \alpha - 1}}$$ for existing class $\boldsymbol{k}$

$$\boldsymbol{p(z = K + 1|Z, \alpha) = CRP(z = K + 1|c, \alpha) = \frac{\alpha}{C + \alpha - 1}}$$ for new class

Infinite Mixtures

- The infinite mixture model can be thought of as the extension of the finite mixture model to the infinite case

$$p(x_i|\theta) = \sum_{k=1}^{\infty} p(x_i|\theta_k)p(z_i = k)$$

$$p(x_i|\theta) = \sum_{k=1}^{\infty} p(x_i|\theta_k)\pi_k$$

- In practice, the 'infinite' here means that it is not fixed in advance.
- So, the number of clusters can grow or shrink.

Finite mixtures : Integrating out π_k (review)

- Using the Dirichlet-Categorical distribution we can integrate out the mixture proportions

$$\begin{aligned} p(X, Z|\theta, \alpha) &= \int p(X, Z, \pi|\theta, \alpha) d\pi \\ &= p(X|Z, \theta) \int p(Z|\pi) p(\pi|\alpha) d\pi \\ &= p(X|Z, \theta) \mathrm{p}(\mathrm{Z}|\alpha) \\ &= p(Z|\alpha) \prod_i p(x_i|Z, \theta) \\ &= p(Z|\alpha) \prod_i p(x_i|\theta_{z_i}) \end{aligned}$$

So: $p(x_i, z_i|Z_{-i}, X_i, \theta, \alpha) = \frac{p(X, Z|\theta, \alpha)}{p(X_{-i}, Z_{-i}|\theta, \alpha)}$

$$= \frac{p(X|Z, \theta)}{p(X_{-i}|Z_{-i}, \theta)} \frac{p(Z|\alpha)}{p(Z_{-i}|\alpha)} = p(x_i|\theta_{z_i})\, p(z_i|Z_{-i}, \alpha)$$

Extension to DP mixtures

- The above equations will remain the same for infinite mixtures.
- The only difference will be that to compute $\boldsymbol{p(z_i|Z_{-i}, \alpha)}$ we will use CRP instead of the posterior conditional Dirichlet categorical
- The CRP is just a special case of the posterior conditional distribution of the Dirichlet categorical
- So:

$$\boldsymbol{p(x_i, z_i = k|Z_{-i}, X_{-i}, \theta, \alpha) = p(x_i|\theta_{z_i})\, p(z_i = k|Z_{-i}, \alpha)}$$

- Since, $\boldsymbol{Z}$, is just a count vector from the perspective of the Dirichlet Categorical

$$\boldsymbol{= p(x_i|\theta_{z_i})\, p(z_i = k|c, \alpha)}$$
$$\boldsymbol{= p(x_i|\theta_{z_i})\, CRP(z_i = k|c, \alpha)}$$

- Thus:

$$\boldsymbol{p(x_i, z_i = k|Z_{-i}, X_{-i}, \theta, \alpha) = p(x_i|\theta_{z_i}) \begin{cases} \frac{c_k}{C + \alpha - 1} & \text{if } k \text{ is an existing class} \\ \frac{\alpha}{C + \alpha - 1} & \text{if } k \text{ is a new class} \end{cases}}$$

- If $\boldsymbol{\theta}$ **is unknown** we sample $\boldsymbol{\theta_k}$ using the posterior distribution:

$$\boldsymbol{p(\theta_k|X,Z,\beta) = p(\theta_k|\{x_i|z_i = k \in Z\},\beta_k)}$$

and then sample:

$$\boldsymbol{p(x_i,z_i = k|Z_{-i},X_{-i},\beta,\alpha) = p(x_i|\theta_k)} \begin{cases} \dfrac{c_k}{C+\alpha-1} & \textit{if k is an existing class} \\ \dfrac{\alpha}{C+\alpha-1} & \textit{if k is a new class} \end{cases}$$

- **Alternatively**, we can **integrate out** $\boldsymbol{\theta_k}$ if $\boldsymbol{\beta_k}$ is a *conjugate prior*
- This will give us a fully collapsed sampler

$$\boldsymbol{p(x_i|X_{-i},z_i = k,\beta) = p\left(x_i\middle|\{x_j|z_j = k \in Z\},\beta_k\right)}$$

$$\boldsymbol{= \int p(\theta_k|\{x_i|z_i = k \in Z\},\beta_k)\, p(x_i|\theta_k) d\theta_k}$$

- In this case, $\boldsymbol{z_i = k}$ can be sampled using:

$$\boldsymbol{p(x_i,z_i = k|Z_{-i},X_{-i},\beta,\alpha) =}$$

$$\boldsymbol{p(x_i|X_{-i},z_i = k)} \begin{cases} \dfrac{c_k}{C+\alpha-1} & \textit{if k is an existing class} \\ \dfrac{\alpha}{C+\alpha-1} & \textit{if k is a new class} \end{cases}$$

Gibbs sampler for infinite mixtures

- Randomly assign $\boldsymbol{z_i = k}$ for some max. K with uniform probability
- Repeat for each $\boldsymbol{i}$:
  - Remove data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$
  - For each cluster $\boldsymbol{k}$ compute $\boldsymbol{p_k(x_i|\,\theta_k)}$ if $\boldsymbol{\theta_k}$ known or can be sampled; alternatively if conjugate, integrate out $\boldsymbol{\theta_k}$ and compute $\boldsymbol{p_k(x_i|\{x_i \in X\,|z_i = k\})}$
  - Add data point $(\boldsymbol{x_i}, \boldsymbol{z_i})$ back by sampling using:

$$\boldsymbol{z_i = k \sim CRP(z_i = k \mid Z_{-i}, \alpha)\, p_k(x_i|\,\theta_k)}$$

- Mixture proportions can be estimated at the end

Discussion

- What we have derived is a DP mixture, i.e., an infinite mixture model
- But we have not actually explained what a Dirichlet process is, since we have only been using the posterior conditional distribution of the DP categorical
- This is like talking about the Dirichlet Categorical distribution while only using the equations for the posterior conditional of the Dirichlet categorical
- So what does the DP look like?
- And can we generate samples from the DP directly?
- A different question is: Is the DP exchangeable? Does it matter if we swap the order in which we sample data points in a different order?
- We will try to answer these.

Exchangeability

Definition: A joint distribution $\boldsymbol{p(x_1, \ldots, x_n)}$ is ***exchangeable*** if for any permutation $\boldsymbol{\pi(x_1, \ldots, x_n) = (x_{\pi_1}, \ldots, x_{\pi_n})}$:

$$\boldsymbol{p(x_1, \ldots, x_n) = p(x_{\pi_1}, \ldots, x_{\pi_n})}$$

- Clearly, if i.i.d. then:

  $\boldsymbol{p(x_1, \ldots, x_n) = p(x_{\pi_1}, \ldots, x_{\pi_n}) = p(x_1) \ldots p(x_n)}$

  hence exchangeable

- But does exchangeability imply i.i.d ?

Exchangeability

Definition: A joint distribution $\boldsymbol{p(x_1, \ldots, x_n)}$ is ***exchangeable*** if for any permutation $\boldsymbol{\pi(x_1, \ldots, x_n) = (x_{\pi_1}, \ldots, x_{\pi_n})}$:

$$\boldsymbol{p(x_1, \ldots, x_n) = p(x_{\pi_1}, \ldots, x_{\pi_n})}$$

- Clearly, if i.i.d. then:

$$\boldsymbol{p(x_1, \ldots, x_n) = p(x_{\pi_1}, \ldots, x_{\pi_n}) = p(x_1) \ldots p(x_n)}$$

hence exchangeable

- But does exchangeability imply i.i.d ?
- Answer: **NO**
- We will show that the Dirichlet process is exchangeable but not i.i.d.

- Lets assume that we are inserting 3 a's into the first cluster and 2 b's into the second cluster. Then:

$$\boldsymbol{p(a,a,a,b,b|\alpha)}$$
$$\boldsymbol{= p(a|\alpha)\, p(a|a,\alpha) p(a|a,a,\alpha) p(b|a,a,a,\alpha) p(b|a,a,a,b,\alpha)}$$
$$\boldsymbol{= \frac{\alpha}{\alpha} \; \frac{1}{1+\alpha} \; \frac{2}{2+\alpha} \; \frac{\alpha}{3+\alpha} \; \frac{1}{4+\alpha}}$$

- Changing the order, we get:

$$\boldsymbol{p(a,b,a,a,b|\alpha)}$$
$$\boldsymbol{= p(a|\alpha)\, p(b|a,\alpha) p(a|a,b,\alpha) p(a|a,b,a,\alpha) p(b|a,b,a,a,\alpha)}$$
$$\boldsymbol{= \frac{\alpha}{\alpha} \; \frac{\alpha}{1+\alpha} \; \frac{1}{2+\alpha} \; \frac{2}{3+\alpha} \; \frac{1}{4+\alpha}}$$

- Hence, the DP is exchangeable and clearly not i.i.d.
- Why is exchangeability important?

- Lets assume that we are inserting 3 a's into the first cluster and 2 b's into the second cluster. Then:

$$\boldsymbol{p(a,a,a,b,b|\alpha)}$$
$$\boldsymbol{= p(a|\alpha)\, p(a|a,\alpha)p(a|a,a,\alpha)p(b|a,a,a,\alpha)p(b|a,a,a,b,\alpha)}$$
$$\boldsymbol{= \frac{\alpha}{\alpha}\ \frac{1}{1+\alpha}\ \frac{2}{2+\alpha}\ \frac{\alpha}{3+\alpha}\ \frac{1}{4+\alpha}}$$

- Changing the order, we get:

$$\boldsymbol{p(a,b,a,a,b|\alpha)}$$
$$\boldsymbol{= p(a|\alpha)\, p(b|a,\alpha)p(a|a,b,\alpha)p(a|a,b,a,\alpha)p(b|a,b,a,a,\alpha)}$$
$$\boldsymbol{= \frac{\alpha}{\alpha}\ \frac{\alpha}{1+\alpha}\ \frac{1}{2+\alpha}\ \frac{2}{3+\alpha}\ \frac{1}{4+\alpha}}$$

- Hence, the DP is exchangeable and clearly not i.i.d.
- Why is exchangeability important?
- Without exchangeability sampling algorithms will give different solutions depending upon order.

The DP Categorical distribution

- Since the CRP is the posterior conditional distribution of the DP categorical, we can work out the full distribution using the same method as in the previous slide.
- Let $\boldsymbol{c} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_k})$ be a vector of counts, then, using the CRP:

$$\boldsymbol{p(c|\alpha)} = \frac{\boldsymbol{\alpha^k}}{\boldsymbol{(1-1+\alpha)(2-1+\alpha)\ldots(C-1+\alpha)}} \prod_{\boldsymbol{i}} \boldsymbol{(c_i-1)!}$$

$$= \frac{\boldsymbol{\alpha^k}}{\boldsymbol{\big(1+(\alpha-1)\big)\big(2+(\alpha-1)\big)\ldots\big(C+(\alpha-1)\big)}} \prod_{\boldsymbol{i}} \boldsymbol{(c_i-1)!}$$

$$= \frac{\boldsymbol{(\alpha-1)!\quad \alpha^k}}{\boldsymbol{\big(C+(\alpha-1)\big)!}} \prod_{\boldsymbol{i}} \boldsymbol{(c_i-1)!} = \frac{\boldsymbol{(\alpha-1)!\quad \alpha^k}}{\boldsymbol{(C+\alpha-1)!}} \prod \boldsymbol{(c_i-1)!}$$

$$= \frac{\boldsymbol{\Gamma(\alpha)\quad \alpha^k}}{\boldsymbol{\Gamma(C+\alpha)}} \prod_{\boldsymbol{i}} \boldsymbol{\Gamma(c_i)}$$

Contrast this with the Dirichlet Categorical:

$$\boldsymbol{p(c|\alpha)} = \frac{\boldsymbol{\Gamma(A)}}{\boldsymbol{\prod_i \Gamma(\alpha_i)}} \frac{\boldsymbol{\prod_i \Gamma(c_i+\alpha_i)}}{\boldsymbol{\Gamma(C+A)}}$$

Dirichlet Process

- The DP is the infinite extension of the Dirichlet distribution
- So, the question is: ***How do we generate samples from the DP***?
- The Stick breaking process is a constructive method for generating samples from the DP.

Stick breaking

- Take a stick of unit length
- Break it into $\boldsymbol{u_1}$ and $\boldsymbol{(1-u_1)}$
- The first portion is $\boldsymbol{u_1}$**.** Call this $\boldsymbol{\theta_1}$ i.e. let $\boldsymbol{\theta_1 = u_1}$
- Break $\boldsymbol{(1-u_1)}$ into $\boldsymbol{u_2}$ and $\boldsymbol{(1-u_2)}$
- The second portion is $\boldsymbol{u_2(1-u_1) = \theta_2}$
  and remainder $\boldsymbol{(1-u_2)(1-u_1)}$
- **….**
- The **k**$^{\textbf{th}}$ portion is

$$\boldsymbol{\theta_k = u_k \prod_{i=1}^{k-1}(1-u_i) = u_k\left(1 - \sum_{i=1}^{k-1}\theta_i\right)}$$

- with remainder

$$\boldsymbol{(1-u_k)\prod_{i=1}^{k-1}(1-u_i) = \prod_{i=1}^{k}(1-u_i) = 1 - \sum_{i=1}^{k}\theta_i}$$

Generating DP using Stick breaking

- Sample $\boldsymbol{u_k \sim Beta(1, \alpha)}$ (simulates breaking stick of unit length)
- then:

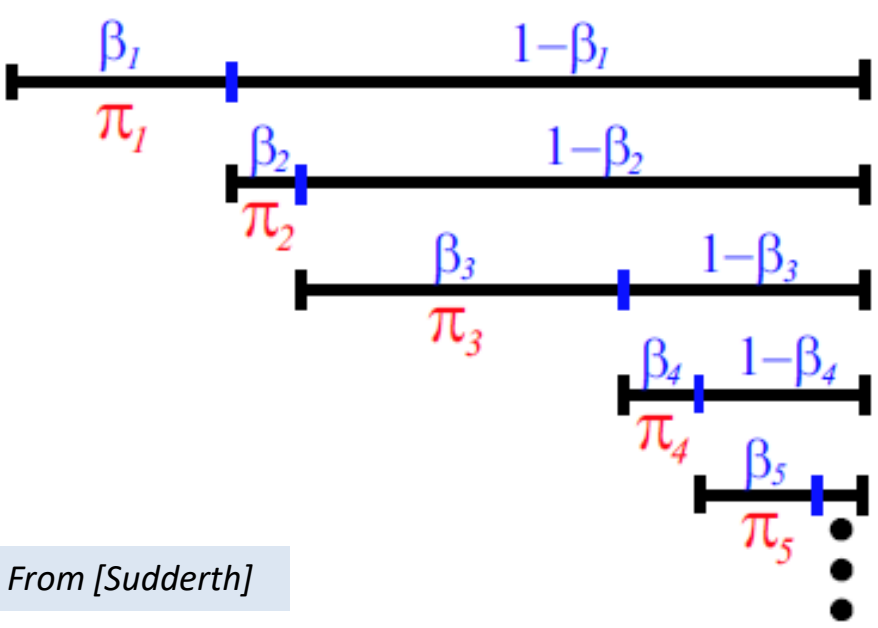
$$\boldsymbol{\theta_k = u_k \prod_{i=1}^{k-1}(1 - u_i) = u_k \left(1 - \sum_{i=1}^{k-1} \theta_i \right)}$$

with remainder:

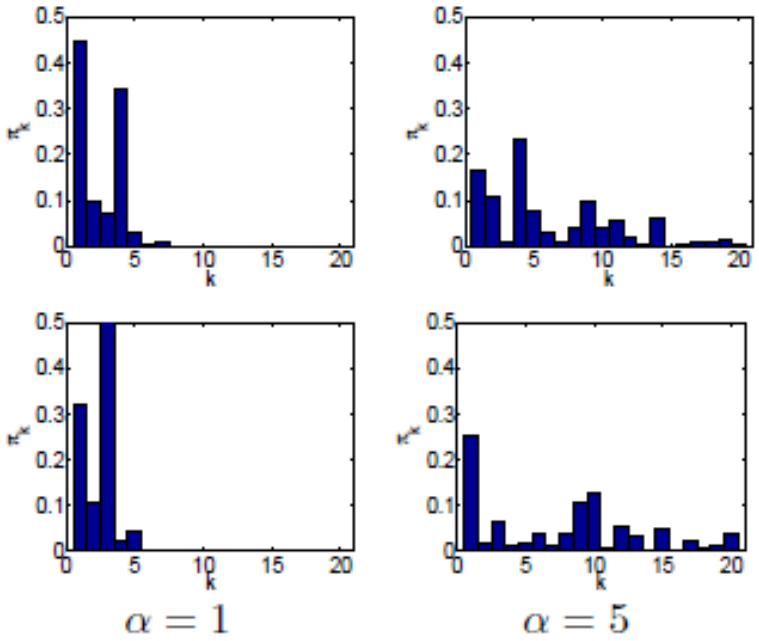
$$\boldsymbol{\prod_{i=1}^{k}(1 - u_i) = 1 - \sum_{i=1}^{k} \theta_i}$$

- Since, the remainder is non-zero, we can go ad infinitum.
- Similar procedure for the Dirichlet distribution exists.

Example sample



From [Sudderth]



- Smaller values of α would result in very dense clusters
- Larger values of α would result in clusters that are more spread out

Hierarchical DP (HDP)

- The problem with standard mixture models is that each mixture component is independent of other mixture components.
- What if some information needs to be shared between mixture components?
- We already looked at an example of this when we used global codebooks for morphology learning.
- What we have so far is a scheme where the mixture proportions are generated by the CRP

- In a hierarchical CRP we assume that individual clusters have their own DPs but will borrow their cluster ids or the value of the z variable from their parent DPs.
- In a two stage DP, we will have one single parent DP.
- Lets assume that the parent DP has counts $\boldsymbol{p} = (\boldsymbol{p_1}, \boldsymbol{p_2}, \dots, \boldsymbol{p_L})$, with $\boldsymbol{P - 1} = \sum_i \boldsymbol{p_i}$
- The counts $\boldsymbol{p_i}$ indicates the number of different tables (i.e. mixture components) sharing the same cluster $\boldsymbol{i}$:

$$\begin{aligned} \boldsymbol{p}(\boldsymbol{x_i}, \boldsymbol{z_i} | \boldsymbol{Z_{-i}}, \boldsymbol{X_{-i}}, \boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{\alpha}) &\propto \boldsymbol{p}(\boldsymbol{x_i} | \boldsymbol{\theta_{z_i}})\, \boldsymbol{p}(\boldsymbol{z_i} | \boldsymbol{Z_{-i}}, \boldsymbol{p}, \boldsymbol{\alpha}) \\ &\propto \boldsymbol{p}(\boldsymbol{x_i} | \boldsymbol{\theta_{z_i}})\, \boldsymbol{p}(\boldsymbol{z_i} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) \\ &\propto \boldsymbol{p}(\boldsymbol{x_i} | \boldsymbol{\theta_{z_i}})\, \boldsymbol{hCRP}(\boldsymbol{z_i} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) \end{aligned}$$

$$\mathbf{h}\boldsymbol{CRP}(\boldsymbol{z_i} = \boldsymbol{k} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) = \frac{\boldsymbol{c_k}}{\boldsymbol{C} + \boldsymbol{\alpha} - \boldsymbol{1}}$$ for existing class $\boldsymbol{k}$ in $\boldsymbol{c}$

$$\mathbf{h}\boldsymbol{CRP}(\boldsymbol{z_i} = \boldsymbol{k} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}}{\boldsymbol{C} + \boldsymbol{\alpha} - \boldsymbol{1}}\ \boldsymbol{CRP}(\boldsymbol{z_i} = \boldsymbol{k} | \boldsymbol{p}, \boldsymbol{\alpha})$$

- In a hierarchical CRP we assume that individual clusters have their own DPs but will borrow their cluster ids or the value of the z variable from their parent DPs.
- In a two stage DP, we will have one single parent DP.
- Lets assume that the parent DP has counts $\boldsymbol{p} = (\boldsymbol{p_1}, \boldsymbol{p_2}, \dots, \boldsymbol{p_L})$, with $\boldsymbol{P - 1} = \sum_i \boldsymbol{p_i}$
- The counts $\boldsymbol{p_i}$ indicates the number of different tables (i.e. mixture components) sharing the same cluster $\boldsymbol{i}$:

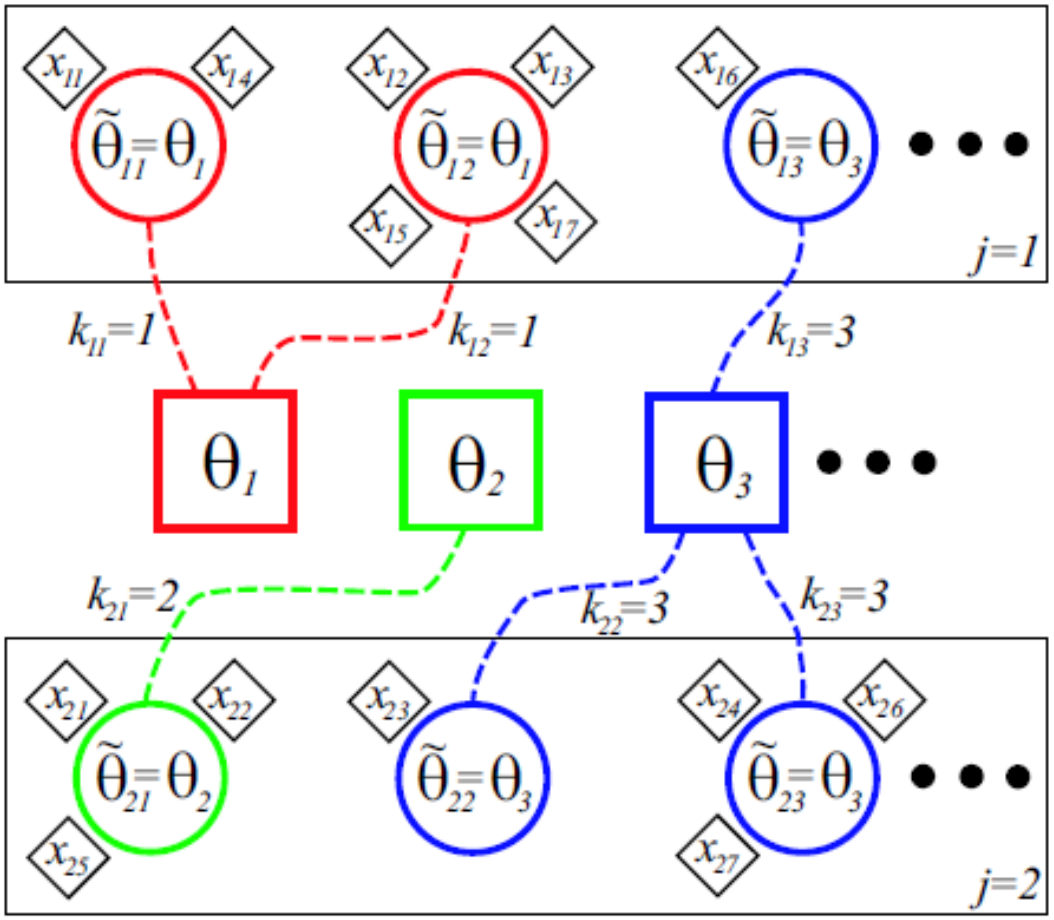
$$\begin{aligned} \boldsymbol{p}(\boldsymbol{x_i}, \boldsymbol{z_i} | \boldsymbol{Z_{-i}}, \boldsymbol{X_{-i}}, \boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{\alpha}) &\propto \boldsymbol{p}(\boldsymbol{x_i} | \boldsymbol{\theta_{z_i}})\, \boldsymbol{p}(\boldsymbol{z_i} | \boldsymbol{Z_{-i}}, \boldsymbol{p}, \boldsymbol{\alpha}) \\ &\propto \boldsymbol{p}(\boldsymbol{x_i} | \boldsymbol{\theta_{z_i}})\, \boldsymbol{p}(\boldsymbol{z_i} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) \\ &\propto \boldsymbol{p}(\boldsymbol{x_i} | \boldsymbol{\theta_{z_i}})\, \boldsymbol{hCRP}(\boldsymbol{z_i} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) \end{aligned}$$

$$\text{h}\boldsymbol{CRP}(\boldsymbol{z_i} = \boldsymbol{k} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) = \frac{\boldsymbol{c_k}}{\boldsymbol{C + \alpha - 1}} \quad \text{for existing class } \boldsymbol{k} \text{ in } \boldsymbol{c}$$

$$\text{h}\boldsymbol{CRP}(\boldsymbol{z_i} = \boldsymbol{k} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}}{\boldsymbol{C + \alpha - 1}} \frac{\boldsymbol{p_k}}{\boldsymbol{P + \beta - 1}} \quad \text{for existing class } \boldsymbol{p_k} \text{ in } \boldsymbol{p}$$

$$\text{h}\boldsymbol{CRP}(\boldsymbol{z_i} = \boldsymbol{L + 1} | \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}}{\boldsymbol{C + \alpha - 1}} \frac{\boldsymbol{\beta}}{\boldsymbol{P + \beta - 1}} \quad \text{for new class } \boldsymbol{L + 1} \text{ in } \boldsymbol{p}$$

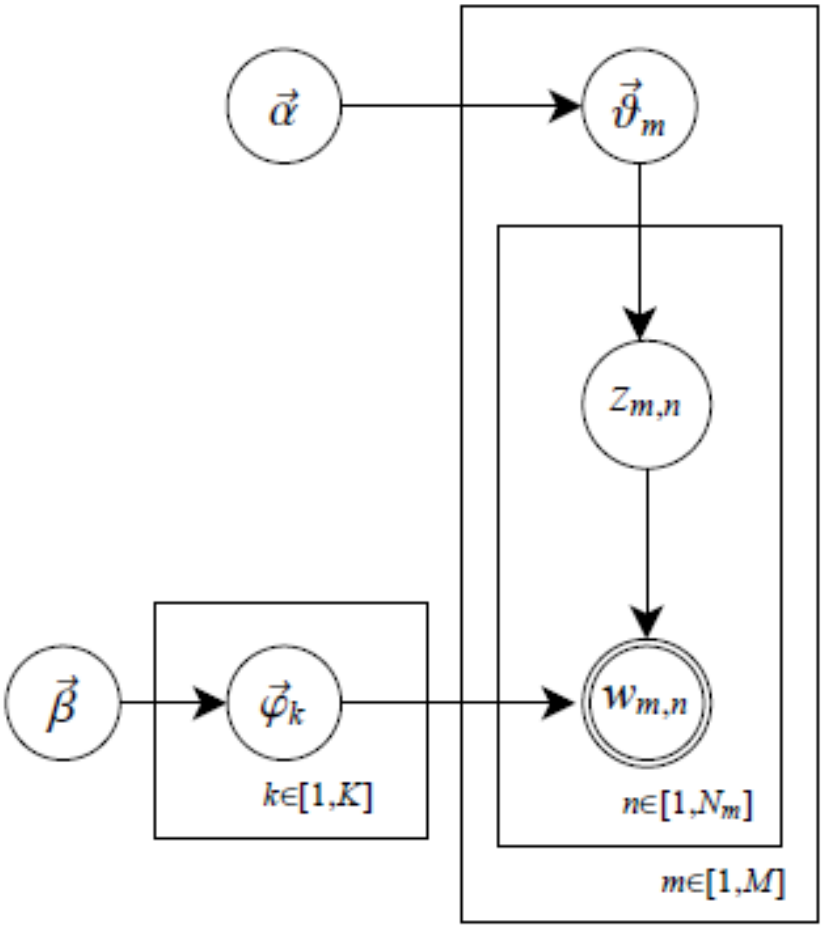
HDP illustration



From [Sudderth]

- Multiple tables in Local DPs may share the same table ids (i.e. dishes)

LDA plate diagram



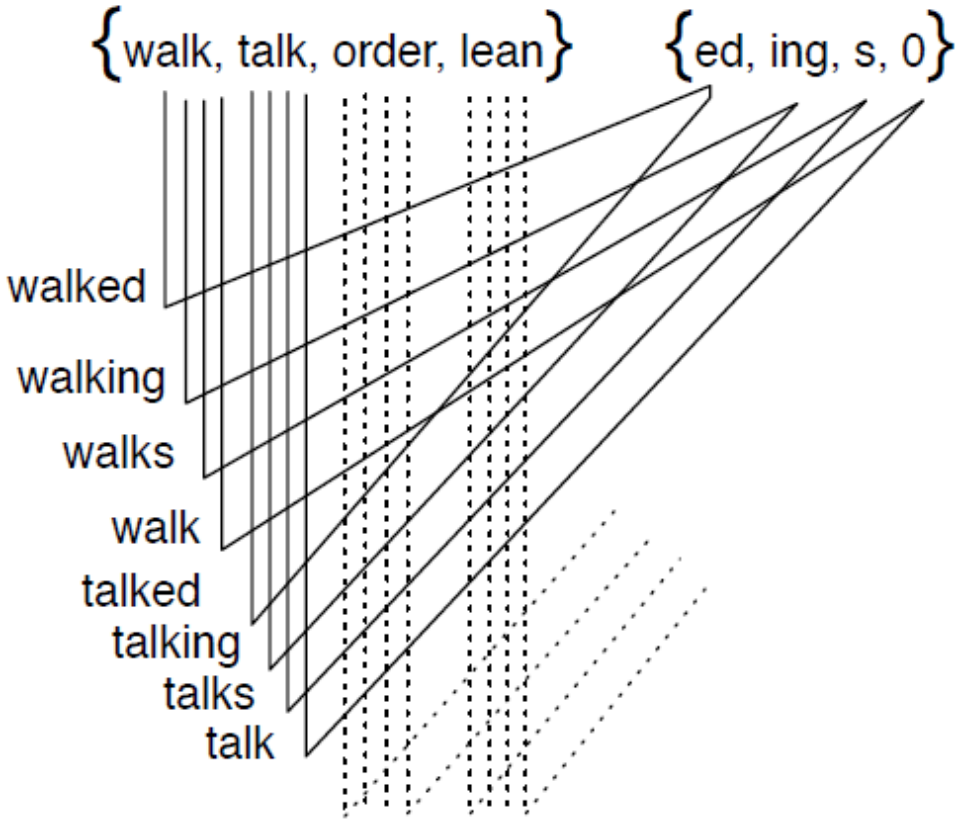
- Use of Dirichlet Multinomials means that the multinomial parameters are integrated out.

Non-parametric LDA

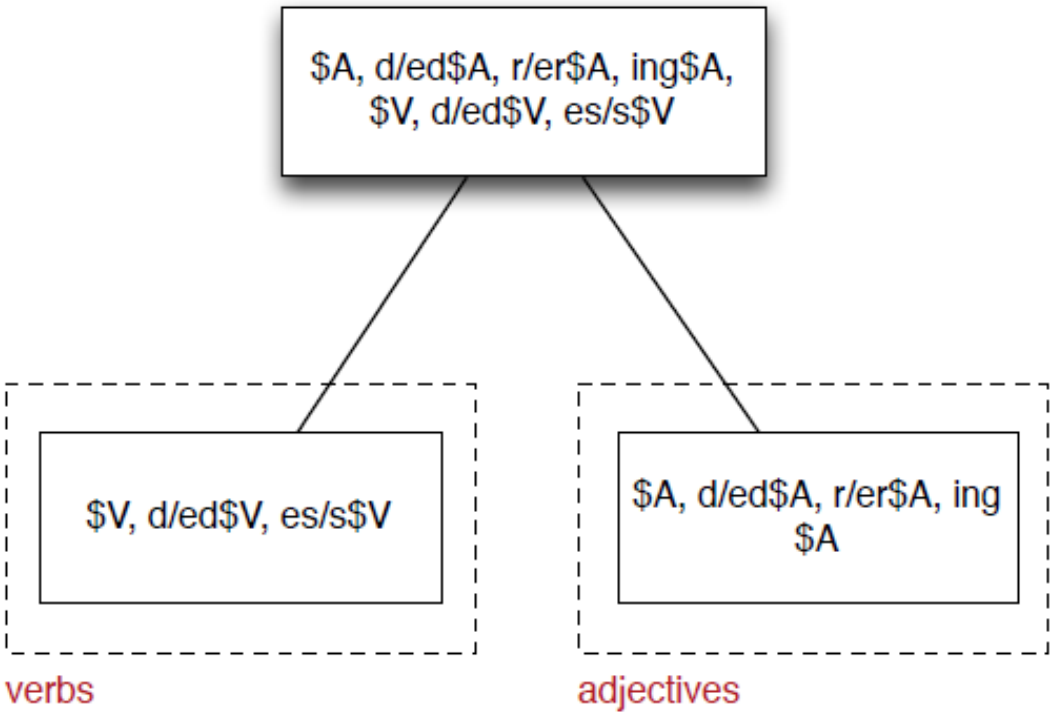
- For LDA where the number of topics need **not** be specified.
- We can employ hCRP scheme
- However, words in a LDA are generated jointly from topic distribution and topic indicators (i.e. the ***z***'s).
- So, we need to ensure that every time a new topic is created, not only should the topic be available for all documents, but topic-word distribution is also shared and updated

- A simple solution is to assume that the Topic-Word distribution is an infinite vector of Dirichlet
- Thus, $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \dots, \boldsymbol{\beta}_\infty)$ where each $\boldsymbol{\beta_i} = \left(\boldsymbol{\beta}_{i_1}, \boldsymbol{\beta}_{i_2}, \dots, \boldsymbol{\beta}_{i_V}\right)$ where $\boldsymbol{V}$ is the length of the vocabulary.
- Each of the $\boldsymbol{\beta_i}$'s are independent from each other
- So, we let each document's $\boldsymbol{Z}$ vector to be generated by hCRP
- Each time a new topic is created within a document, hCRP ensures that the topic id is shared across all documents
- Now, since the $\boldsymbol{\beta}$ vector is infinite, for every (globally) new topic $\boldsymbol{i}$ there is a readily available parameter vector $\boldsymbol{\beta_i}$ that generates the words.
- $\boldsymbol{\beta}$ needs to be implemented e.g. as a hash table since, as topics die and are re-born not all topic ids will survive
- For example, there could be topics {1, 25, 47, 50}
- To manage this efficiently, keep a list of ***dead topic ids*** for reuse

Learning Hierarchical Morphology



Burcu Can and S. Manandhar. *Probabilistic Hierarchical Clustering of Morphological Paradigms*, EACL 2012.

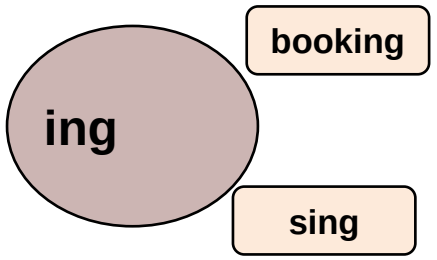


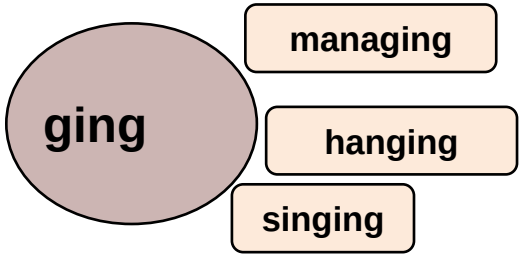
- Key idea is to learn morphological paradigms and arrange them in a hierarchy

Independent DPs for Stems/Suffixes

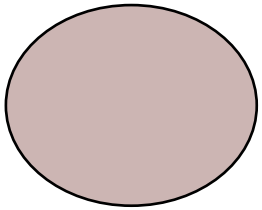
binging *new word*

suffix restaurant





empty table



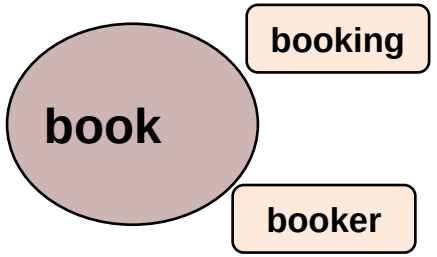
$p(ing) = 2/(N-1+ \beta) * p(ing)$

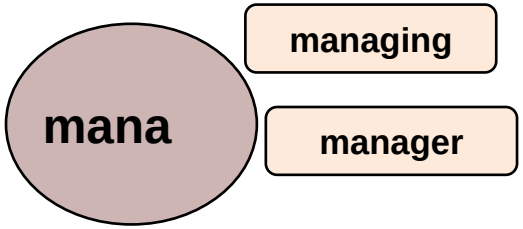
$p(ging) = 3/(N-1+ \beta) * p(ging)$

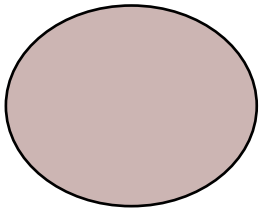
$p(ng) = \beta/(N-1+\beta) * p(ng)$

---

stem restaurant

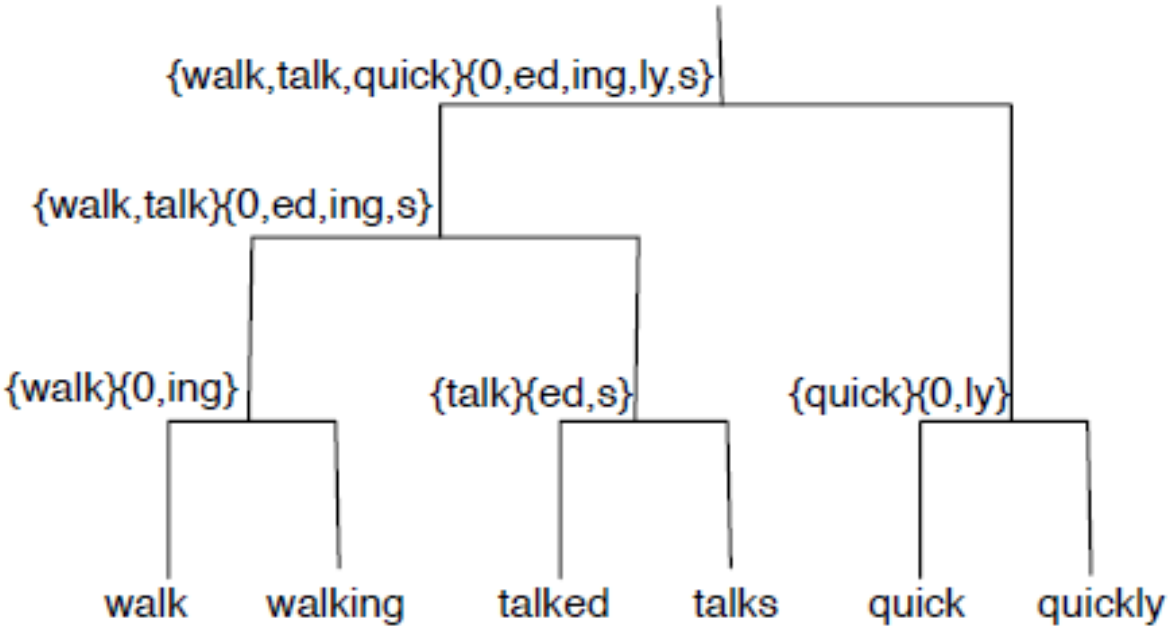






Data covered by a tree node

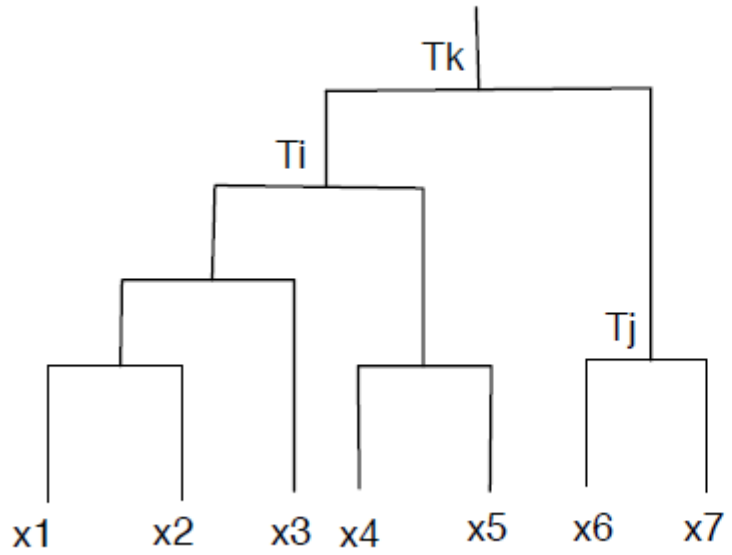
- Dataset: $\mathbf{D} = \{w_1 = s_1 + m_1, \ldots, w_n = s_n + m_n\}$



- Data points covered by a tree node are all words dominated by that node

Likelihood of a tree

- Dataset: $\mathbf{D} = \{x_1, x_2, \ldots, x_n\}$. T is the entire tree.



$$p(D_k|T_k) = p(D_k)p(D_i|T_i)p(D_j|T_j)$$

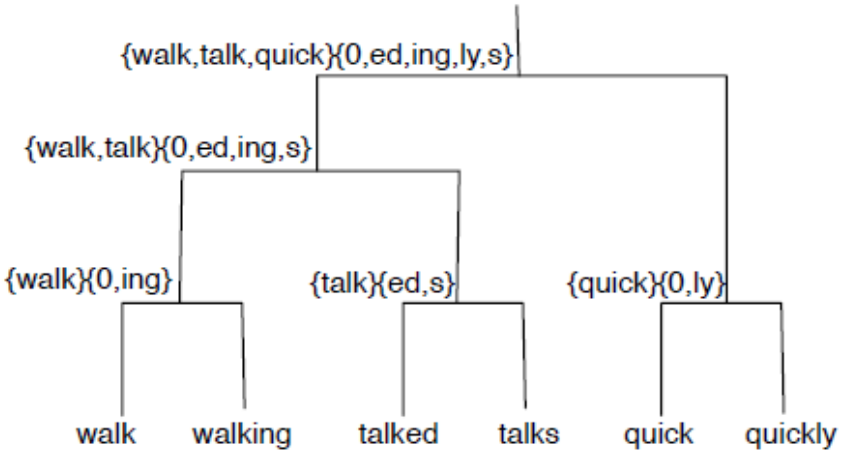
$$\mathbf{D_i} = \{x_1, x_2, x_3, x_4, x_5\}$$

$$\mathbf{D_j} = \{x_6, x_7\}$$

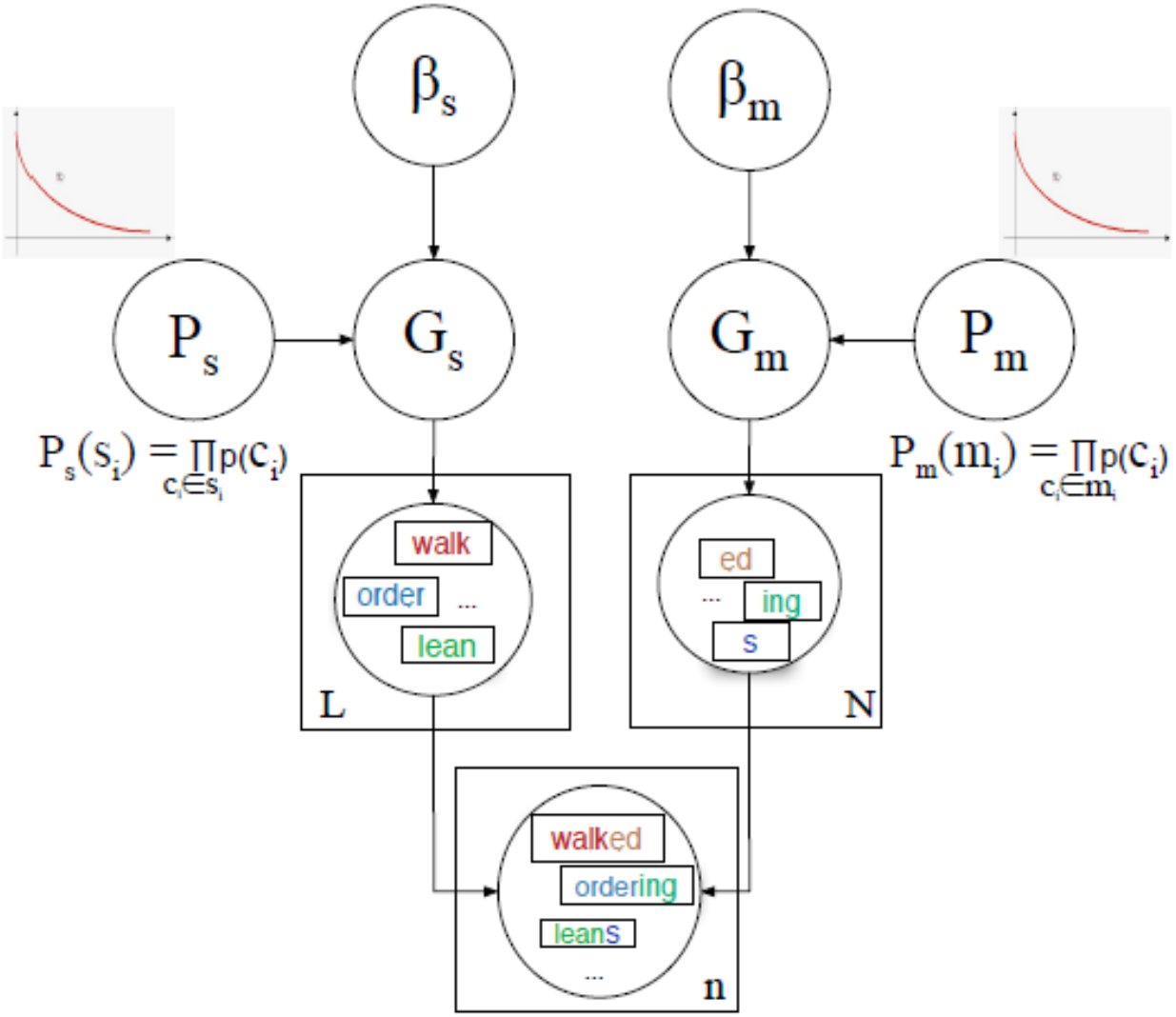
- Remember : Likelihood means $\boldsymbol{p(X|\theta)}$ i.e. the probability of data given the model parameters.
- Since, $\boldsymbol{\theta}$ is a tree node $\boldsymbol{k}$ here. $\boldsymbol{\theta}$ becomes $\boldsymbol{T_k}$.
- And since, data covered by tree node is $\boldsymbol{D_k}$ we write $\boldsymbol{p(D_k|T_k)}$

Local likelihood

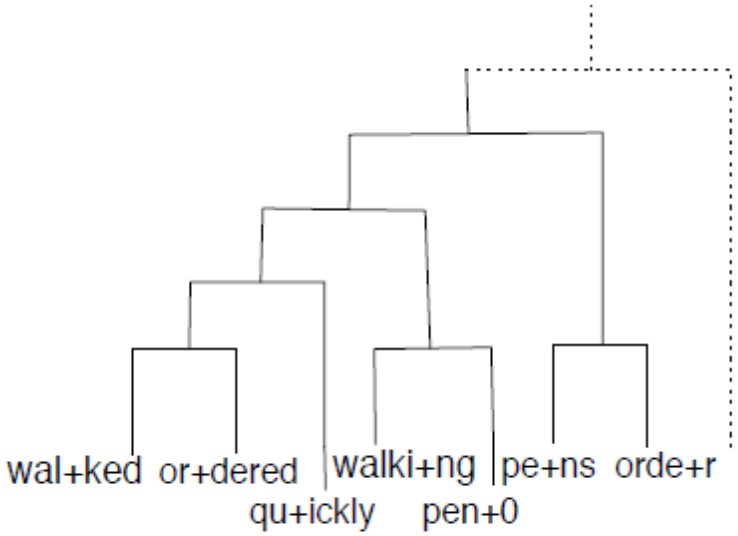
- Dataset: $\mathbf{D} = \{w_1 = s_1 + m_1, \ldots, w_n = s_n + m_n\}$



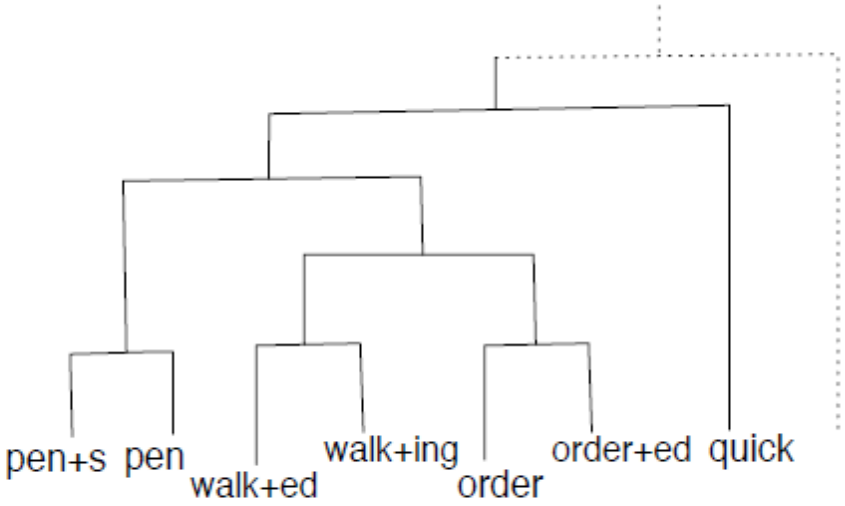
$$p(D_k) = p(S_k)p(M_k)$$
$$= p(s_1, s_2, \ldots, s_n)p(m_1, m_2, \ldots, m_n)$$


β_s
β_m
P_s
G_s
G_m
P_m
$P_s(s_i) = \prod_{c_i \in s_i} p(c_i)$
$P_m(m_i) = \prod_{c_i \in m_i} p(c_i)$
walk
order
...
lean
L
ed
...
ing
s
N
walked
ordering
leans
...
n

Construct initial random tree



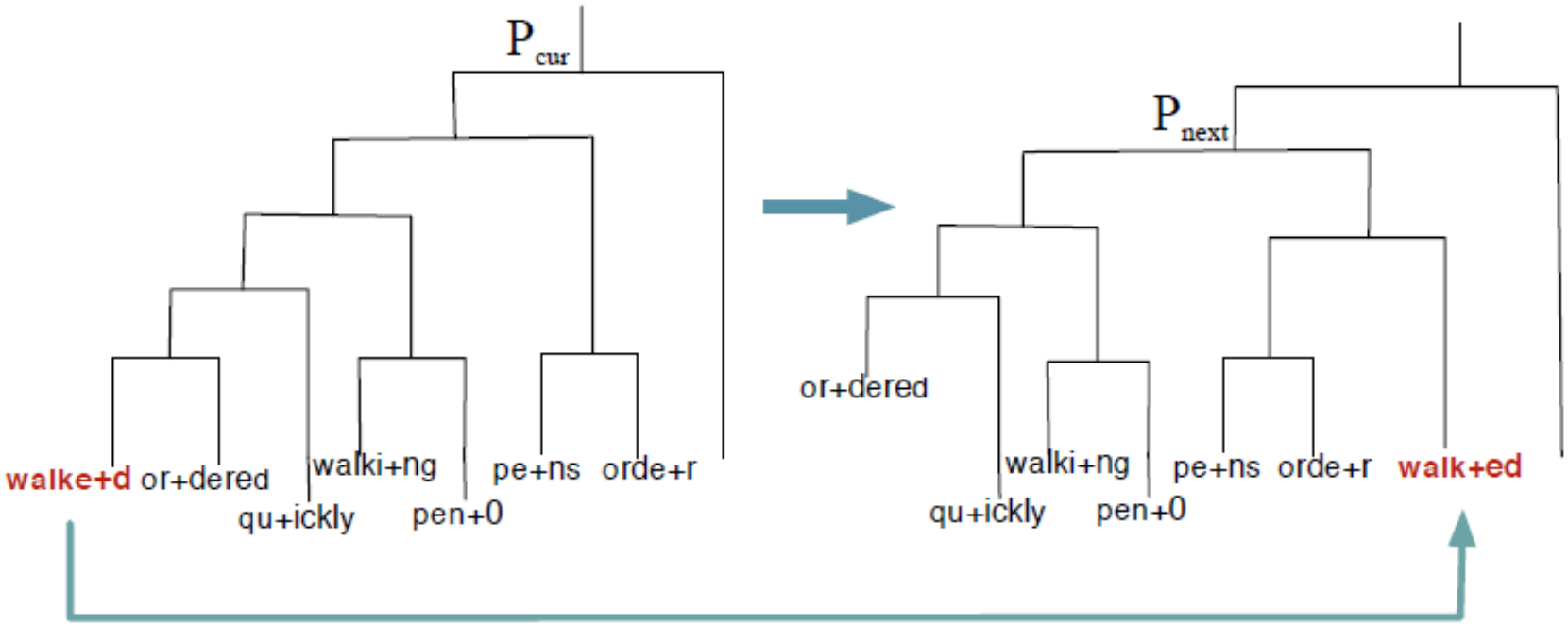
1) Construct the initial tree
(randomly generated)



2) Apply inference
(final tree)

- Generate a randomly created initial tree
- Move to the next tree configuration using sampling steps

Metropolis Hastings sampling



Accept if $p_{next}(D|T) > p_{cur}(D|T)$

else accept with probability $P_{Acc} = \frac{p_{next}(D|T)}{p_{cur}(D|T)}$

- Gibbs sampling can be thought of as a special case of MH.
- In Gibbs, the 'if' part does not exist

Computing Likelihood

- The marginal likelihood of words in the node k is defined such that:

$$p(D_k) = p(s_1, s_2, \ldots, s_n)p(m_1, m_2, \ldots, m_n)$$

- The joint probability of stems $S = \{s_1, \ldots, s_L\}$ is:

$$\begin{aligned} & p(s_1, s_2, \ldots, s_L) \\ & = p(s_1)p(s_2|s_1)\ldots p(s_M|s_1, \ldots, s_{M-1}) \\ & = \frac{\Gamma(\beta_s)}{\Gamma(L+\beta_s)}\beta_s^K \prod_{i=1}^{K} P_s(s_i) \prod_{i=1}^{K}(n_{s_i}-1)! \end{aligned}$$

- The joint probability of suffixes $M = \{m_1, \ldots, m_N\}$ is:

$$\begin{aligned} & p(m_1, m_2, \ldots, m_N) \\ & = p(m_1)p(m_2|m_1)\ldots p(m_N|m_1, \ldots, m_{N-1}) \\ & = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\alpha^T \prod_{i=1}^{T} P_m(m_i) \prod_{i=1}^{T}(n_{m_i}-1)! \end{aligned}$$

T

- Since, MH based on using likelihoods and not conditional likelihoods as in Gibbs, we need to compute the full likelihood.
- So, this is just DP categorical distribution as worked out in slide 24

References

- **Tutorials**

Peter Orbanz. *Lecture Notes on Bayesian Nonparametrics*. Columbia University

Erik B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD Thesis, MIT, 2006.

Yee Whye Teh. *Introduction to Bayesian Nonparametrics*. Machine Learning Summer School (MLSS) 2011.

- **Papers on Dirichlet Processes**

Various papers by Radford Neal on Dirichlet Processes and sampling schemes

- **Hierarchical Morphology clustering based on the following paper**

Burcu Can and Suresh Manandhar. Probabilistic Hierarchical Clustering Of Morphological Paradigms. In *13th Conference of the European Chapter of the Association for computational Linguistics* (EACL), 2012.